Investigating Demographic Features and their Connection to Performance, Predictions, and Fairness in EDM Models

Lea Cohausz
University of Mannheim
lea.cohausz@uni-mannheim.de

Andrej Tschalzev
University of Mannheim
andrej.tschalzev@uni-mannheim.de

Christian Bartelt
University of Mannheim
christian.bartelt@uni-mannheim.de

Heiner Stuckenschmidt
University of Mannheim
heiner.stuckenschmidt@uni-mannheim.de

Although using demographic features for predictive models in Educational Data Mining (EDM) has to be considered very problematic from a fairness point of view and is currently critically discussed in the field, they are, in practice, frequently used without much deliberate thought. Their use and the discussion around their use mostly rely on the belief that they help achieve high model performance. In this paper, we theoretically and empirically assess the mechanisms that make them relevant for prediction and what this means for notions of fairness. Using four datasets for at-risk prediction, we find evidence that removing demographic features does not usually lead to a decrease in performance but also that we may sometimes be wrong in aiming to achieve the most accurate predictions. Furthermore, we show that models, nonetheless, place weight on these features when they are included – highlighting the need to exclude them. Additionally, we show that even when demographic features are excluded, some fairness concerns relating to group fairness metrics may persist. These findings strongly highlight the need to know more about the causal mechanisms underlying the data and to think critically about demographic features in each specific setting – emphasizing the need for more research on how demographic features influence educational attainment. Our code is available at: `https://github.com/atschalz/edm`.

Keywords: demographic features, sensitive features, fairness, algorithmic bias, at-risk prediction

1. Introduction

One of the recent discussions in Educational Data Mining (EDM) concerns the question of whether demographic features should be used for predictive models (Baker et al., 2023). Demographic features "refer to particular characteristics of a population [. . .], such as age, race, gender, ethnicity, religion, income, education, [. . .]" (Salkind, 2010). Baker et al. (2023) provide a definition of demographic features in the context of Educational Data Mining (EDM) specifically, stating that they encode "information that provides context about students' backgrounds and experiences that they bring with them to school" and that (also outside of the context of education) "demographic variables cannot be manipulated".

Whether demographic features should be included in predictive models is connected to the fairness notion of "fairness through unawareness" (Castelnovo et al., 2022; Dwork et al., 2012). The concept of fairness through unawareness has been discussed in the literature of fairness in Machine Learning (ML) for a while (Mehrabi et al., 2021; Caton and Haas, 2024), and has recently garnered attention in EDM as well (Baker et al., 2023; Yu et al., 2021; Kizilcec and Lee, 2022). The key idea behind this fairness notion is that predictive models that place importance on demographic features are not fair because people who are otherwise equal may get different predictions as a result (Castelnovo et al., 2022; Kizilcec and Lee, 2022). For example, suppose we use a model to admit students to a course based on their prediction to finish the course. If a demographic variable impacts this prediction, people who are equally well-suited according to all other features may get a different prediction solely based on demographic characteristics. Clearly, this happening is something that most would deem unfair, and many advocate excluding demographic features to avoid it (Hu and Rangwala, 2020; Baker et al., 2023).

This position is not necessarily shared by every researcher. Yu et al. (2021) state that demographic features should be included and that not doing so can be "construed as subscribing to a 'colorblind' ideology" (p. 98). In particular, Baker et al. (2023) argued theoretically against the position of Yu et al. (2021) and counter that one should exclude demographic features in most settings due to concerns for biases. Most importantly, a model trained on biased labels may reinforce these biases, and models may not be as effective for small groups when demographic features are used[1]. In practice, however, it has been observed that many studies in EDM use demographic features for their predictive models (Alturki et al., 2022; Paquette et al., 2020; Baker et al., 2023). Alturki et al. (2022) evaluated the features most used across EDM studies predicting student success from 2007-2018. Among the ten most used features are six demographic features (gender, age, income, nationality, marital status, employment status) – the most common of which is gender. The implicit and sometimes explicit reasoning behind both the argument brought forward by Yu et al. (2021) as well as the frequent use of demographic features, is that these features matter in the context of education.

It is, of course, the case that the demographic background influences educational attainment drastically. Affluent students and students with well-educated parents perform better (Azhar et al., 2014), and women are seen as less competent compared to their equally well-performing male counterparts in the classroom setting (Bloodhart et al., 2020). These are just two examples of demographic characteristics influencing educational attainment. They also highlight why excluding them from predictive models is important – otherwise and as stated by Baker et al. (2023) the biased data leads to biased models – but investigating their effect outside of using them for predictive models is interesting and worthwhile.

Although the importance of demographic features for educational attainment is well known, their impact on predictions in at-risk EDM models is uncertain. Some studies indicate that using them influences the model's performance (Batool et al., 2021; Daud et al., 2017), whereas other studies indicate the opposite (Tomasevic et al., 2020; Jha et al., 2019). These mixed results led Kizilcec and Lee to state that "it is unclear whether the inclusion of protected attributes adds predictive value to algorithmic systems, and the merits of fairness through unawareness may be mostly symbolic" (p. 13) (Kizilcec and Lee, 2022).

If, however, demographic features do not add anything to predictive models regarding the performance of those models, then there can be two reasons for this. Either demographic fea-

[1] In addition, Baker et al. (2023) also argue against their use because it may not lead to actionable insights.

tures do not matter at all and have no value for the prediction (which is unlikely given that their importance is generally known), or all information that is encoded in them is also encoded in the other predictive features, in which case new questions about fairness arise, namely whether it is fair that those other features encode demographic information (Cohausz et al., 2024; Kizilcec and Lee, 2022). Because the other features encode demographic information, the average predictions for different demographic groups will still vary when fairness through unawareness is employed. The notion of fairness through unawareness is known for this shortcoming (Kizilcec and Lee, 2022). Other fairness notions, most importantly and prominently, group fairness notions, can capture this emerging fairness problem (Kizilcec and Lee, 2022; Castelnovo et al., 2022).

So far, however, a detailed discussion and evaluation of the role of demographic features considering performance and fairness is missing. Therefore, this paper will investigate the effects of using and not using demographic features for prediction across four publicly available EDM datasets that focus on at-risk prediction. In doing so, we will investigate the effects of demographic features regarding both model performance and fairness. In detail, the contributions of this paper are the following:

- We provide a discussion of causal mechanisms that may lead to demographic features being used in predictive models and relate this to fairness concerns and notions.
- We subsequently evaluate how demographic features are used in predictive models and arrive at the conclusion that we should, in most cases, remove demographic features as they will be used by models otherwise. We also highlight that removing demographic features does not lead to a decrease in model performance, which may be an additional incentive to remove them, but it also leads to the question of whether we should even aim for the best predictive performance. Even when removing the demographic features, the influence of demographic features may persist through other features encoding information about them, and, on occasion, removing them might even make the models more biased.

By doing so, we extend the study by Cohausz et al. (2023) that already showed that including demographic features generally does not lead to a performance boost but did not embed this in a fairness context. Note that our paper only focuses on at-risk and student success prediction. While demographic features are also used in other areas of EDM, we believe it is better to focus on one subset of tasks for now and thoroughly investigate this (Paquette et al., 2020). Furthermore, note again that we are only concerned with the use of demographic features in predictive models in this study, not with how exactly demographic features influence educational attainment. Investigating how and why demographic features matter should, of course, be encouraged.

2. Theoretical Considerations on Demographic Features and Fairness

2.1. Types of Features

Before discussing the impact of demographic features, we want to briefly define and state which features exist in EDM at-risk predictions and how we define the categories. In accordance with Tomasevic et al. (2020), we argue that there are three major types of features in EDM:

demographic, performance, and activity/engagement features. Additionally, we identify features not belonging to any of the categories.

Demographic Features. Demographic features are traditionally considered to be features that refer to characteristics of a population and are not changeable within the educational context immediately (Baker et al., 2023). Typically used demographic features are gender, age, ethnicity, nationality, or features indicating socioeconomic status, e.g., parental occupations or household income. Furthermore, we need to consider the existence of proxy variables. As Baker et al. (2023) write, a proxy variable "correlates sufficiently highly with a demographic variable that using the proxy variable produces results (and bias) similar to using the demographic variable (perhaps to a lesser degree)" (p. 39). Note, however, that variables may also exist that are highly correlated with the demographic feature but that are not, according to Baker et al. (2023), proxy variables because they are also on their own correlated with the target, i.e., they are mediators that are on the path between the demographic feature and the target. The distinction between proxy variables and mediating variables is a difficult one, however, because we usually do not know the exact causal relationships involved. While for some features, it may be clear that they can only be proxy variables (e.g., the zip code is a proxy for socioeconomic status and perhaps race), this distinction may be less certain for other variables. In general, if we know that certain variables are proxy variables, we should treat them in the same way as demographic features, whereas mediating variables may be used differently (as we will discuss later). Furthermore, note that the definition of demographic features is relatively strict. For example, a feature that encodes a student's amount of free time (which might be influenced by factors such as having a part-time job, which is in turn influenced by socio-demographic aspects) will not be counted as a demographic feature as the school can heavily influence it with its timetable and extracurricular activities. Additionally, the way in which demographic characteristics influence education may vary depending on the setting. For example, in some countries, female students may be disadvantaged while receiving equal treatment to their male counterparts in other countries. Given that we evaluate the effect of demographic features across four datasets and focus on the general influence of demographic features for predictions (not for individual features' specific contributions), we cannot go into detail for every dataset but point out variables where the classification may be contentious.

Performance Features. Any study-related performance measures, e.g., grades, information on passes or fails, or percentages on assignments, are considered performance features. In other words, any information that hints at how well a student did in the past belongs to this type.

Activity Features. Activity features are features that are study-related and show how active a student is. Typical features of this type are, among others, participation during class, hours spent on online learning platforms, and participation in online forums. This category also includes extracurricular activities as long as they are performed in school.

Other Features. Most features in EDM datasets belong to one of the above categories. Other features not belonging to either of these categories would be, e.g., the study program, the semester a student is in, or when the course takes place. We also placed features in this category that could be argued to be demographic features (e.g., free time, alcohol consumption) if we employed a different definition.

As our focus is on investigating whether using demographic features is advantageous when we also have some study-related features, we do not differentiate between activity and performance data. For the remainder of the paper, we define study-related features as all features related to a student's study activity and previous performance.

2.2. Fairness Notions

As mentioned in Section 1, demographic features evoke fairness concerns. When we speak of fairness in Machine Learning (ML), we typically speak about algorithmic bias, i.e., about models that lead to differing results for specific demographic groups (Mehrabi et al., 2021).

Rather obviously, this may happen if demographic features are assigned importance in models and, hence, one fairness notion is that these features should be removed: this is the notion of fairness through unawareness as mentioned in the introduction (Castelnovo et al., 2022). As also already mentioned, however, a model that conforms to fairness through unawareness does not necessarily conform to other notions of fairness. Other notions of fairness encompass individual fairness, causal fairness, and, most prominently, statistical group fairness measures. Individual fairness relies on the idea that similar people, according to some distance metric, should receive similar predictions. Not only is it a challenge for this fairness notion to define an appropriate distance metric, but the notion additionally still has the same problem that fairness through unawareness also has: if other features are correlated with the demographic information, it will still be the case that, on average, one demographic group has a smaller or higher probability to receive a specific prediction (Castelnovo et al., 2022; Kizilcec and Lee, 2022).

Fairness concerns that become evident when considering the groups of people of specific characteristics are captured by group fairness metrics. Specifically, when we want to consider that the predictions still differ for different groups even when removing demographic features, we are interested in the notion of independence (Castelnovo et al., 2022; Baker et al., 2023; Kizilcec and Lee, 2022).

The group fairness metric that encodes the idea of independence is Demographic Parity (DP). It states that the selection rate, i.e., the probability that the positive label is predicted, should be equal for each demographic group (Bird et al., 2021; Mehrabi et al., 2021). This notion can be expressed as:

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b)$$

Where $\hat{Y}$ is the prediction, and A is the demographic feature. Note that DP does not rely on the real target values at all; it makes no assumptions about whether the target is biased or unbiased. This means that a biased model will be considered unfair when looking at DP, even if this bias is also reflected in the true distribution of the target values. If we want to make sure that our model is fair according to DP, we need to mitigate the bias using a bias mitigation strategy that ensures that DP is met (Deho et al., 2022). However, whether we believe that a lack of independence, i.e., DP indicating a bias, is problematic is a subjective and situation-dependent view – this is generally true for all group fairness notions.

Another group fairness metric tightly connected to DP is that of conditional independence expressed by the Conditional Demographic Parity (CDP) metric (Castelnovo et al., 2022). It posits that the selection rate should not differ conditional on other legitimate features X:

$$P(\hat{Y} = 1|A = 0, X = x) = P(\hat{Y} = 1|A = 1, X = x), \forall x$$

Interestingly, fairness through unawareness and CDP may indicate a fair or unfair model in the same cases: if we declare the legitimate features X to be all features that are not demographic, both notions call all models fair if the models do not directly rely on demographic features. Of course, we may decide that only certain other features are legitimate, in which case the two notions can lead to different results. Like DP, CDP also makes no assumptions about the target.

This lack of assumption about the target is not the case for all group fairness metrics. Suppose we assume that the differences with regard to the predictions are fully justified by the actual differences among demographic groups with regard to the true label. In that case, we should turn to other metrics, such as predictive equality (equal false positive rates), equality of opportunity (equal false negative rates), and equalized odds (the combination of both) (Castelnovo et al., 2022). Each of these metrics is concerned with achieving equally accurate models for all demographic groups. They are not concerned with an equal distribution of the target classes across all demographic groups. Therefore, if we remove the demographic features, these metrics should not be affected as much as the general accuracy for each group will decrease or remain stable.

Hence, when we talk about the consequences of removing demographic features, we will focus solely on DP as this metric can be affected and can show whether the fairness problems, according to this notion, persist when removing demographic features. Still, we want to note again that just because DP indicates a problem does not mean that this is a problem in all cases and according to every person who judges the fairness of models (Cohausz et al., 2024).

For the sake of completion, we also want to briefly mention causal fairness. For causal fairness to be in place, we need to make sure that a person's prediction would not change if they belonged to a different demographic group (Castelnovo et al., 2022; Kizilcec and Lee, 2022). As group membership also affects other features, we would need to know the exact causal mechanisms to judge whether a change in group membership would change the prediction. As we usually do not know the exact causal mechanisms, using this fairness notion is not feasible (Kizilcec and Lee, 2022).

2.3. The Role of Demographic Features

We now discuss why demographic features may be used in models and, hence, the sources of algorithmic bias according to DP. Algorithmic bias usually follows from models placing importance on demographic features. The importance stems from them being involved causally in the generation of the target, which means that they are usually correlated with the target and, therefore, relevant for prediction (Cohausz et al., 2024). Hence, discussing how they are involved causally in the data generation mechanism means discussing both their impact on the predictive performance and the algorithmic bias of a model. Cohausz et al. (2024) recently showed different causal relationships involving demographic features. A causal relationship between a feature X and a feature Y means that X causes Y and is denoted as $X \rightarrow Y$. Hence, these relationships show how the data is generated. We will now discuss the causal relationships and how they relate to performance and bias. As already said, note that in reality, we hardly ever know how exactly features influence other features causally. The scenarios below showcase how relationships could look and how they relate to DP.

The causal relationship between demographic features and the target may be a direct one, i.e., $A \rightarrow Y$ for demographic feature A and target Y (Cohausz et al., 2024). For example, women are perceived as less able even if they are achieving equally good or even better grades than their male counterparts (Bloodhart et al., 2020). If women and men apply to graduate school, it is easy to imagine that gender directly influences the admission decision. In this case, removing the demographic features achieves fairness according to DP, as the information on the demographic features is not encoded elsewhere. Furthermore, removing demographic features leads to a drop in performance. However, this can also be viewed as desirable because the target is clearly also biased (e.g., more men are truly admitted, which leads to a biased target distribution).

The relationship may also be indirect, i.e., $A \rightarrow X \rightarrow Y$ where X is some other predictive feature that acts as a mediator (Cohausz et al., 2024; Baker et al., 2023). For example, Hill et al. (2010) found that women receive letters of recommendation with remarks that are viewed less positively in comparison to those that men receive. As an example of these remarks, women were more likely to receive recommendations that stress their compassion, whereas men received more praise for their ability and scientific contribution (Hill et al., 2010). If we again think about using the recommendation letter to predict admission, it is clear that gender will, at least indirectly, influence the admission, as the recommendation letter also encodes this information (Pearl, 2009). Removing demographic features will not mean that DP is achieved, and likewise, the model's performance will not or will only slightly drop, showcasing the limits of the notion of "fairness through unawareness." In such situations, it should be questioned whether we would still consider the model fair and whether some kind of bias mitigation needs to be performed on X.

Of course, it is also possible that both indirect and direct causal relationships apply simultaneously (e.g., women are perceived as less competent, and their recommendation letters contain less favorable attributes), in which case the DP only indicates a fairer but not fair model, and the performance decreases a bit. The question of how to handle X then still remains. Furthermore, it is also possible that the demographic feature is a confounder (i.e., a cause for both) for another predictive variable X and the target, i.e., $X \leftarrow A \rightarrow Y$. In such a case, we may call X a proxy variable of A, and if we know that the causal relationship looks like this, we should remove it (Baker et al., 2023). If we do not remove X, importance will be placed on it, resulting in the model still being biased and probably only a minor dip in performance.

For all of the cases discussed so far, the true target is also biased. As already stated, such a setting means that our goal should no longer be to receive maximally accurate models. There are, though, also situations where the target is not biased. Such a situation, where the target is not biased, but the ML model might be, occurs when a demographic feature influences another feature X that shares a known or unknown confounder with the target, i.e., $A \rightarrow X \leftarrow C \rightarrow Y$, where C in the confounder (Cohausz et al., 2024). In this case, because of the correlation with Y, X might be used to predict the unbiased target, but it will make the prediction biased because X is influenced by the demographic feature. Curiously, here, it could even be the case that removing the demographic feature increases the bias according to DP (Cohausz et al., 2024). If we include it, then it might correct the bias from the other predictive feature as doing so makes the prediction more correct – meaning that DP will not indicate a bias. But if we exclude it, the correction is not possible, and DP will indicate a bias. What follows is that for this scenario, fairness through unawareness may lead to models that one may consider not fair. Accordingly, the performance of the model will drop if we remove the demographic feature in the case where the confounder is not known because X is biased, whereas the target is not. It will not drop if the confounder is known as the confounder will then be used. In this scenario, because the target is not biased, we should feel comfortable with trying to achieve a highly accurate model.

A final source of algorithmic bias may be underrepresentation. If only very few people in our data have a specific demographic characteristic and all of them have the same value for the target by chance, the model may overfit to this characteristic, even though other features can explain the target (Mehrabi et al., 2021). Removing the feature will then still indicate a bias – even though it does not really exist (Cohausz et al., 2024). It will not affect the predictive performance, though, as the relevant information can be taken from other features.

To summarize our theoretical considerations, demographic features can impact the prediction in a multitude of ways. If we do not only have a direct connection, removing the demographic features will not mitigate the bias concerns, according to DP, and it will also not necessarily lead to a drop in performance. Removing the features is still usually a very good idea as the models might otherwise use them. Removing demographic features is not necessarily the right strategy when we have a confounder, though, that impacts both the target, which is unbiased, and another predictive feature, which is influenced by a demographic feature. In this case, removing the demographic feature negatively influences both performance and DP. This latter scenario is also the only scenario for which we can certainly say that we aim for a maximally accurate model; in the other cases, the target itself is biased, which should at least evoke the question of whether we really should aim for highly predictive models. Regardless of the exact scenario, it is always possible that ML models place weight on demographic features.

Before we turn to our empirical evaluation of the effects of using and not using demographic features that we theorized about, we first want to consider existing research that focused on the usefulness of demographic features for predictions to receive an idea of how much they matter for performance. A detailed empirical evaluation of their impact on algorithmic bias according to DP does not yet exist.

3. Existing Evidence

Existing empirical research is divided on whether demographic features are useful for accurate predictions.

3.1. Demographic Features Are Useful for Predictive Performance

Batool et al. (2021) used the popular Open University Learning Analytics Dataset (OULAD) and two similarly structured datasets and selected only the demographic features in the datasets to predict who will fail the courses. They report high F1 scores using Random Forests but do not compare against baselines to validate the meaningfulness of their results. Daud et al. (2017) predict whether a student will finish their degree based on socio-demographic features using a dataset from several universities in Pakistan. They considered many features not typically available and potentially extremely problematic, e.g., family expenditures. Daud et al. (2017) report high F1 scores but do not compare this against predictions using previous performance data. Hoffait and Schyns (2017) predict which students are at risk at the time of registration for their degree using a dataset from Belgium. Due to their setting, they only have some previous performance data from school and no activity information, but most of their data is demographic. Yet, they achieve relatively high F1 scores.

3.2. Demographic Features Are Not Useful for Predictive Performance

Tomasevic et al. (2020) also used the OULAD to predict performance and compared several machine learning models with different sets (demographic, performance, activity) of features against each other in a very thorough study. Usually, the prediction accuracy did not vary much when using or not using demographic features as long as the other sets of study-related features were used, leading them to conclude that these features were not important, although using demographic features usually slightly improved the model. At least for this dataset, this finding is very strong evidence that demographic features do not significantly add to the prediction accuracy. They apparently did not use all demographic features available. Al-Zawqari and Vandersteen (2022) used a subset of the OULAD dataset to distinguish between high-performing and failing students. They compared F1-scores using and not using demographic data along with activity data and found that using demographic data did not improve results much. It should be noted that it is unclear how they selected and handled their data. Jha et al. (2019) used the same dataset to predict failure using a variety of methods and different feature subsets. In accordance with the other papers, they found that activity data was the most predictive feature set. When they used activity data, it did not matter what other features were included regarding the model's performance. Trstenjak and Donko (2014) used data from the Information System of Higher Education Institutions databases to predict success and rank feature importance using several metrics such as information gain and gain ratio. They showed that most (but not all) demographic features had very little impact and experimented with leaving some (the least important ones) of them out, which even led to slightly increased accuracy. Miguéis et al. (2018) predicted the overall study success of students of a technical university and then looked at the Gini-index of features. They found that performance data was more important than demographic data.

3.3. Demographic Features Are Somewhat Useful for Predictive Performance

Khasanah et al. (2017) predicted overall study success with data from Indonesia. They used information gain to evaluate demographic feature importance and found that some were important, but others were not. It should be noted that the data they had available on previous performance and activity was rather limited. Sweeney et al. (2016) looked at the feature importance of one large dataset as they tried to predict study success for the courses a student enrolled in the next term. They found that demographic data were more important in the beginning when little past performance data was available than later on. However, they had relatively few demographic features in their dataset. Zhao et al. (2020) used admissions data to predict who will perform well in a specific Master's program based on admission data. Due to the nature of their setting – that they try to learn who should be admitted to the program – their performance data is restricted to data on high school and Bachelor results, and they have no activity data. Though they make no distinction between demographic and non-demographic features, their most important predictors show that some demographic features (gender, nationality) tend to be important while others are not. Cortez and Silva (2008) predicted grades of Portuguese middle school students in math and Portuguese. They found that the relative importance of previous performance scores was higher, but socio-demographic features still mattered.

3.4. Overall Evidence

Overall, for the case of OULAD, despite the results of Batool et al. (2021), the evidence appears to be clear that accuracy does not increase when using demographic data along with performance or activity data (Tomasevic et al., 2020; Jha et al., 2019). In general, studies that included study-related features typically found demographic features to be less important. However, in other settings with less performance data, results suggest that demographic data plays a role. Those who explicitly investigated feature importance typically reported that it is somewhat important. Furthermore, note that only very few studies explicitly reported on feature engineering of demographic characteristics. Yet, feature engineering is often non-trivial for demographic data as it often consists of (high-cardinality) categorical data.

In accordance with our theoretical considerations, the existing evidence seems to support the following conclusions: demographic features do impact the predictions as they do lead to good results when other data is not available. So it is generally, and as expected, not the case that they are simply unimportant. But, when having other data, the features are no longer important when it comes to increasing predictive performance. Yet, when looking at the feature importance scores, some importance is still given to them even when other features are included. Both observations indicate that an indirect relationship is at play or that the demographic features act as a confounder and also influence other features that are, therefore, correlated with the target.

4. Research Objective and Questions

We now want to investigate the effects of using and not using demographic features on model performance and fairness. Both our review of existing evidence and our theoretical considerations lead us to the hypothesis that using demographic features will not increase model performance as long as we have study-related features from previous performance or activity but that they will have predictive power if we do not have study-related features. **Accordingly**, we hypothesize that even if we do not include demographic features, fairness metrics might still indicate that a bias exists as the information of at least some demographic features might be encoded in the other predictive features.

To test our hypothesis, we formulate the following research questions:

- **Research Question 1 (RQ1)**: Do demographic characteristics explain at least some of the differences in student performance; in other words, do models using only demographic features perform better than guessing? With this research question, we investigate whether demographic features are relevant to the model performance at all, i.e., whether they even transport relevant information. If they are not relevant, we would not even need to remove demographic features as they would not be used. If they are helpful, though, this implies that the target itself is biased.

- **Research Question 2 (RQ2)**: Are demographic characteristics still helpful for model performance if study-related information is available; in other words, do models trained on study-related and demographic features perform better than models trained only on study-related features? With this research question, we test whether all information that the demographic features provide for the target is also completely encoded in other predictive features. If this is the case, removing demographic features does not affect the performance of the model but also does not decrease fairness concerns.

- **Research Question 3 (RQ3)**: If RQ2 is answered with no, do models trained on the whole data learn that demographic information is irrelevant for the prediction; in other words, do models trained on the whole data place close to zero importance on the demographic features? With this research question, we can answer whether removing demographic features is necessary from the perspective of fairness through unawareness. If models place importance on demographic features, we should remove them.

- **Research Question 4 (RQ4)**: If RQ2 is answered with no, does removing demographic features reduce fairness concerns; in other words, does demographic parity vary between models with and without demographic features? With this research question, we test whether the bias, as indicated by DP, decreases, remains stable, or even increases when removing demographic features. The result tells us about the concerns about fairness shown by DP, which remain even when we remove demographic features.

5. Experimental Design

In this section, we describe our experimental setup to evaluate the formulated research questions. We proceed by first describing the datasets and model classes used for prediction. Afterward, the hyperparameter tuning procedure, methods to treat categorical data, and the evaluation setup are described.

5.1. Datasets

We use four publicly available EDM datasets. Two datasets are from online learning systems and two from in-class education; of the latter, one is from secondary education in high schools and one from tertiary university education.

In this subsection, we briefly describe the used datasets and the corresponding preprocessing. Furthermore, we describe the assignment of features to the feature types (demographic, performance-related, activity-related, and others) discussed in Subsection 2.1. We will use the resulting feature subsets in Section 6 to train models for answering the research questions. An overview of the datasets can be seen in Table 1. The summary of which features are allocated to which feature category can be seen in Table 2.

5.1.1. Dataset of Academic Performance Evolution for Engineering Students

The dataset of academic performance evolution for engineering students (Delahoz-Dominguez et al., 2020) consists of the academic, social, and economic information of $12,411$ Columbian engineering students. Student performance was assessed at two points in time: in the final year of high school and in the final year of their professional training in Engineering. We refer to this dataset as *Engineering*. The first assessment evaluates five generic academic competencies:

Table 1: Description of the datasets used to evaluate the research questions.

	Engineering	PortSecStud	xAPI-Edu	OULAD
No. of samples	12411	1044	480	22437
No. of features	33	34	17	51
Performance features	5	3	0	4
Demographic features	25	17	4	6
Activity features	0	6	5	40
Other features	2	7	7	0
Categorical features	13	4	7	4
Total cardinality	3980	17	59	31
% NA	0.0	0	0.0	0.48
Target $\mathbf{y} \in$	[1..166]	[1..19]	[1..3]	$\{1,2\}$

Table 2: Description of the allocations of features to subsets for each dataset.

	Demographic Features	Study-Related Features	Other Features
Engineering	gender; parental, geographic, and school information; item availability in family	first assessment on five dimensions (MAT, CR, CC, BIO, ENG)	university; academic program
PortSecStud	gender; age; address; family and school related information; paid classes; internet access	first and second period grade; past failures, absences; study time; extracurricular activities	lifestyle related features, e.g. alcohol consumption, romantic relationships, amount of free time
xAPI-Edu	gender; nationality; place of birth; parent responsible	interaction with the e-learning system; absences	general academic information (e.g. semester, field of study); parental participation
OULAD	gender; region; imd_band; age_band; disability; highest_education	num_of_prev_attempts; avg_cma; avg_tma; studied_credits; sum of clicks and count of visits for each of the 20 VLE activity types	-

mathematics, critical reading, citizen competencies, biology, and English. The second assessment evaluates critical reading, quantitative reasoning, citizen competencies, written communication, English, and the formulation of engineering projects.

As the target for predictions, we use the global score of the second performance assessment and treat the task as a regression task. The five dimensions of the first assessment are used as performance information. There is no information about student activity in the dataset. Demographic features include gender, parental education and occupation, geographic information, school information, and whether different items, such as a car or computer, were available in the family. Other available information is the university and the academic program a student attends. Note that the university may be correlated with students' socioeconomic background. It is, however, also an interesting predictive feature outside of this (universities may also differ in their teaching style, specialization options, etc.). Additionally, one may transfer to another university for various reasons, which makes this a feature that can be changed within the educational context. Because of this, we leave it in the "other" category, but there may be arguments for placing it in the demographic category, too. The identifier features, as well as all dimensions and variants of the performance assessment besides the global score, are excluded. Further, dataset-specific preprocessing is not necessary. Thirteen categorical features are in the dataset, of which two are of very high cardinality (i.e., more than ten values). There are students from 3,735 schools and 134 universities.

5.1.2. Dataset of Portuguese Secondary School Student Performance

The dataset (Cortez and Silva, 2008) consists of students from secondary education in two Portuguese schools (mean age: 16.7) and can be used to predict student achievement in math and

Portuguese language courses. We refer to this dataset as *PortSecStud* The target is the final course grade, which is measured on a discrete scale between 0 and 20. Some authors categorize the grade into pass and fail for binary classification or into five levels for classification. However, we consider it a regression problem, as it better represents the nature of the problem. As performance information, the first and second-period grades are available, as well as the number of past class failures. Activity information consists of the weekly study time, absences, and whether the student participated in extracurricular activities. The demographic information includes gender, age, and address, as well as school and family-related information. Furthermore, we considered travel time from home to school, educational support from family, extra paid classes within the course subject, and having internet access as demographic features since they are highly influenced by socioeconomic factors and cannot be changed by the school. Other features are lifestyle-related, such as alcohol consumption or whether the student is engaged in a romantic relationship. Note that some may argue that a few of the features in this category could be seen as demographic as well. Because the school might influence these features and they are not usually listed as demographic features, we decided to use them as other features. The datasets for the math and Portuguese courses are combined, and a feature indicating the course has been added. Further, dataset-specific preprocessing is not necessary.

5.1.3. xAPI-Edu-Data

The Students' Academic Performance Dataset (xAPI-Edu-Data) (Amrieh et al., 2016) consists of 480 underage students, where most are from Kuwait (179) and Jordan (172). Data was collected through a learning management system which required internet access. The target is students' performance in %, which is only available in groups: 0-69, 70-89, and 90-100. Hence, we treat the task as a multi-class classification problem. There is no information about previous student performance in the dataset. Student activity is measured according to four behavioral aspects during interactions with the e-learning system: participation in discussion groups, visiting resources, raising a hand in class, and viewing announcements. In addition, absence days are available. Demographic features are nationality, gender, place of birth, and the parent responsible for the student. Other information includes basic academic information (e.g., course, semester, grade level) and the parents' participation (answering a survey, school satisfaction). With this classification, we follow the classification by Amrieh et al. (2016), who published the dataset. No dataset-specific preprocessing is required. The categorical features with the most expressions are nationality, with 14 possible nationalities, and field of study, with 12 possible subjects.

5.1.4. OULAD

The OULAD dataset is a large dataset with diverse opportunities for educational data mining (Kuzilek et al., 2017). The data stems from the participants of online courses that adults with diverse academic and financial backgrounds can take in the United Kingdom. The data is stored in a relational database of five tables with information on students, assessments, courses, registrations, online learning materials, and students' interactions with the materials. We focus on the same prediction task with the same dataset, features, and preprocessing as Jha et al. (2019). For predictions, we consider all students who did not drop out before the course ended to predict whether they failed or passed. As information about the previous performance, we use the average scores achieved in previous assignments. Jha et al. (2019) conducted analyses on different data subsets as well; however, they counted the so-far achieved credits and the number of previous attempts as demographic features. This does not match our definition of demographic features, so we instead define those features as performance-related. Student activity is obtained as two types of interaction with 20 different content types, resulting in 40 features. The types of interaction are the sum of the clicks and the number of visits for each type of content. Content types include homepage, subpage, quiz, wiki, and other platform-related types. As demographic features, we use gender, region, imd_band (an index of deprivation of areas in the UK), age_band, and disability. There are no other features in the dataset. The performance and activity features are extracted from the database as described by Jha et al. (2019). Similarly, the id_student, code_module, module_presentation, and exam_score features were excluded, as well as all students who had withdrawn before the course ended. Some mean assessment scores and imd_band categories are missing. The missing assessment scores are due to there not being an assessment in this course. We accounted for this by adding a value representing this. As the information on how missing values are treated is not given by Jha et al. (2019), we impute the mean value for the mean assessment scores and define a new category for missing imd_band values.

5.2. Models

In our evaluation, we include two model classes, namely generalized linear models (GLMs) and XGBoost. For regression tasks, we use lasso regression for the regularization of the models to prevent overfitting. For classification tasks, we use logistic regression with the L2-penalty. In the case of multi-class classification, multinomial loss is used. GLMs have the benefit of being highly interpretable and, thus, are ideally suited for (educational) data mining. However, they strongly assume that the target's relationship to the features is linear. In contrast, XGBoost is a highly flexible model capable of learning more complex relationships. For the OULAD dataset, XGBoost has been shown to outperform competitive approaches by Jha et al. (2019). Furthermore, for tabular datasets, XGBoost has shown superior performance compared to other methods like neural networks far beyond the field of educational data mining (Shwartz-Ziv and Armon, 2022; Grinsztajn et al., 2022). Thus, it can be considered the state-of-the-art model for maximizing performance on a variety of datasets such that we do not include further models. In addition, a simple baseline for each dataset is included, which predicts the target mean of the training data for regression tasks and the mode for classification tasks. By comparing models trained solely on demographic data to these baselines, we are able to answer research question **RQ1**. For each dataset, we learn models for different feature subsets: using only demographic data, using only study-related data, using study-related data and demographic data, and using all data (i.e., also including the other features). This setup allows us to answer **RQ2**.

5.3. Hyperparameter Optimization

We implement a hyperparameter optimization (HPO) pipeline with 5-fold cross-validation (5CV) for XGBoost and GLMs. For parameter tuning, we use Bayesian optimization implemented in the hyperopt library (Bergstra et al., 2015). We chose Bayesian hyperparameter optimization because the method searches the search space more efficiently, reduces computational cost (in comparison to grid search), and balances exploration and exploitation well (Wu et al., 2019). In practice, it has been found to work well for a variety of ML methods, including tree-based methods (Wu et al., 2019). To select the best parameters the training data is split into five folds again, i.e., we have a nested approach. In each HPO step, a model with the current hyperparameters is trained on each fold. The objective function of each step is the average performance on the held-out datasets of each fold. Our modeling pipeline is depicted in Figure 1. Performance is measured as the mean squared error (MSE) for regression tasks and log-loss for classification tasks. For the GLMs, we only tune the regularization strength parameter α. The search space for Lasso regression is defined as $\alpha \in [10^{-10}, 0.5]$. The search space for logistic regression is defined as $\alpha \in [10^{-10}, 1.0]$. We run 50 iterations of Bayesian optimization for each model.

For hyperparameter optimization of XGBoost, we implement an algorithm to iteratively tune different subsets of XGBoost hyperparameters using Bayesian optimization in four steps.

1. Tune the number of estimators $\in [50..500]$ and the learning rate $\in [0.001, 0.5]$.
2. Tune the maximum tree depth $\in [1..18]$ and minimum child weight $\in [0..10]$.
3. Tune both the number of columns and samples used in each tree $\in [0.5, 1]$.
4. Tune the regularization parameters $\alpha \in [0..10]$, $\lambda \in [1, 4]$ and $\gamma \in [10^{-8}, 9]$.

In each step, 50 iterations of Bayesian optimization are performed. To speed up the computations and terminate the training for optimization iterations with poor parameter choices more quickly, we use early stopping on the validation data if there is no improvement after ten training iterations. Overfitting on the validation data is mitigated through the 5CV procedure as a configuration needs to perform well on all five validation sets.

Note that our hyperparameter optimization might affect our results for **RQ3**, which is concerned with the importance assigned to demographic features. As our hyperparameter optimization chooses a hyperparameter configuration that considers all features, importance might be placed on features that are not necessarily required. To ascertain that our results for **RQ3** are not simply due to the hyperparameter optimization, we also checked the results using the default settings for a sklearn linear model and XGBoost. Doing so either produced the same patterns or was not applicable due to bad performance, showcasing that tuning is necessary. Furthermore, note that other machine learning methods, such as neural networks, would be more prone to place weight on all features (Grinsztajn et al., 2022).

5.4. Methods for Categorical Data Treatment

All of the used datasets include categorical data. As the treatment of categorical data, in particular, high-cardinal data, can affect predictive performance in data mining tasks (Pargent et al., 2022), we first evaluated whether our models are affected by different encoding methods. We tested a total of five different encoding techniques, namely One-Hot-Encoding, Ordinal Encoding, Target Encoding, Catboost Encoding, and Regularized Target Encoding (Micci-Barreca, 2001; Pargent et al., 2022; Prokhorenkova et al., 2018). All of the encoding methods have different advantages and disadvantages depending on the setting. We compared the results using different encoding techniques and found no major differences. Hence, we ultimately decided to use Regularized Target Encoding for the GLM models and Ordinal Encoding for the XGBoost models, as these encodings are specialized for these methods. The detailed results of our evaluation can be seen in Appendix A.

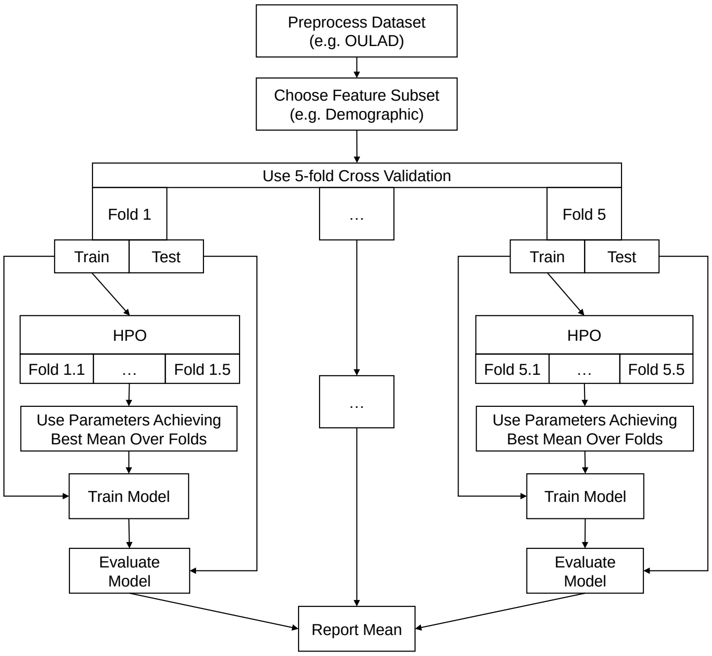

Figure 1: Data pipeline for model development and evaluation.

5.5. Predictive Performance Evaluation

For regression tasks, the target is normalized to zero mean and unit variance for training the models, then the predictions are denormalized afterward to interpret the performance on the original scale. All continuous features were normalized to zero mean and unit variance as well. Doing so allows us to interpret GLM coefficients as feature importance scores directly. For each configuration, we use 5-fold cross-validation (5CV) for evaluation. As an evaluation metric, we use root mean squared error (RMSE) for regression (lower is better) and F1-score with macro averaging for classification (higher is better). We always applied a paired t-test (significance level < 0.05) to check whether the methods performed significantly better.

5.6. Methods for Determining the Impact of Demographic Features in Models

As we investigate the use of demographic data in EDM predictions, performance is not the only relevant metric. It is equally important to analyze the extent to which the models use demographic data. If they are not used, removing them is not necessary. Hence, to answer our research question **RQ3**, we analyze the feature importance of trained models with a focus on demographic data. We analyze the learned coefficients of the linear models as well as the feature importance of the XGBoost models. As we normalized the data, the coefficients of linear models can directly be interpreted as feature importance scores. For the linear models, we first normalize the absolute coefficient values to sum to one. Afterward, we sum the normalized coefficients for the demographic features to obtain an assessment of the extent to which the models use demographic data for predictions. For XGBoost, the feature importances reflect how often certain features are used and how useful they are for the prediction in a single decision tree.

Precisely, the importance of a single tree is calculated as the amount that each split improves the performance measure, weighted by the number of samples the node is responsible for. Afterward, the feature importances are averaged across all of the decision trees and normalized to sum to one.

In addition, we analyze the extent to which the utilization of demographic data affects the actual predictions on linear models; this is considered important information in the fairness literature (Mehrabi et al., 2021). Given a dataset with n samples and d features in a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and a target $\mathbf{y} \in \mathbb{R}^n$, we apply the following procedure:

1. Train a linear model to predict the target
2. Obtain predictions as $\hat{\mathbf{y}} = \sigma(\mathbf{X}\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the coefficient vector of the linear model and σ is the inverse link function depending on the target, e.g., linear for continuous and sigmoid for binary targets
3. Remove the k demographic features from $\mathbf{X}$ and the respective coefficients $\boldsymbol{\beta}$ and obtain predictions $\tilde{\mathbf{y}} = \sigma(\mathbf{X}_{:,d-k}\boldsymbol{\beta}_{d-k})$
4. Compute score for the impact of demographic features as
   (a) $\frac{1}{n}\sum_i^n \hat{y}_i - \tilde{y}_i$ for regression – which means that values can range from 0 to $max(y)$
   (b) $\frac{1}{n}\sum_i^n \{1 \text{ if } \hat{y}_i \neq \tilde{y}_i, 0 \text{ otherwise } \}$ for classification – which means that values can range from 0 to 1

For regression, this corresponds to evaluating the mean absolute difference of predictions with and without demographic features. For classification, this corresponds to evaluating the percentage of samples for which not using the demographic features changes the class assignment.

5.7. Fairness Metrics

In order to assess whether models not using demographic features are potentially still containing biases against certain demographic groups and, thus, answer **RQ4**, we will – as already discussed in Subsection 2.2 – employ the metric of demographic parity (DP). To be more precise, we decided to use the demographic parity ratio, as it allows an immediate assessment of the bias indicated by this metric (Castelnovo et al., 2022). The parity ratio is the ratio of the smallest selection rate and the largest selection rate of any demographic group. As mentioned in Subsection 2.2, the selection rate is the average probability that a member of a certain demographic group is assigned the positive value (Bird et al., 2021). What follows is that a demographic parity ratio value close to 1 indicates that the rates for the different groups are almost similar and most likely no bias exists, whereas a value close to 0 indicates the opposite. In research, a 4/5 rule, which has its origin in legal regulation, is frequently mentioned, meaning that if the value is 0.8 or greater, we consider the model unbiased (Bird et al., 2021). The validity of this assessment does not always hold and depends on the setting, but it is often at least a sensible guideline. Furthermore, we want to stress again that a metric indicating a bias is nothing but a hint that an actual fairness concern could exist – and an absence of such an indication is not a guarantee that some kind of fairness notion is definitely achieved (Cohausz et al., 2024). This statement is particularly true as different people may evaluate fairness differently, as already discussed in Subsection 2.2. For this work, we are only interested in whether there is still an indication of a bias corresponding to the fairness notion most applicable when removing demographic features. Therefore, while we are comfortable using this metric, we will refrain from using the word fairness during the evaluation of **RQ4** and will speak of a bias indication. In order to apply demographic parity to the models, we proceed as follows.

Handling continuous targets. DP is usually only computed for classification and not regression problems. In order to still receive information for the regression problems, we compare the average prediction for each group (akin to the DP) by calculating the ratio. For simplicity, we also refer to this as DP.

Handling extremely underrepresented groups. If a demographic feature contains a value shared by only very few people (e.g., nationality) in absolute numbers (mostly a problem in the xAPI-Edu dataset), DP might indicate strong biases, although the differences between groups are only due to chance (e.g., everyone of a group consisting of three students passed because they just happened to all be good students). However, it is nearly impossible to tell whether the difference is due to chance or not. Hence, if a value has less than 20 instances, we recode it to "other".

Just like for the other evaluation metrics, we computed the values for each of the folds and took the average.[2]

6. Results

In this section, we report and discuss our results to evaluate the stated research questions. We start with a comparison of models trained on different data subsets. Subsequently, we evaluate the feature importance of the models using all data to assess whether demographic data is used. Finally, we evaluate the demographic parity to see whether biases still exist.

6.1. Performance Comparison of Different Feature Subsets

Table 3: The table shows the means and standard deviations resulting from 5CV on different data subsets. Root mean squared error (where lower is better) is reported for Engineering and PortSecStud, and F1-score (where higher is better) for xAPI-Edu and OULAD. Results per column for methods that are not significantly different from the best method in a paired t-test (alpha=0.05) are highlighted in bold.

Dataset		Baseline	Demo only	Study only	Demo + Study	All
Engineering	GLM	23.11 (0.26)	20.53 (0.3)	14.47 (0.22)	14.35 (0.25)	**14.14 (0.29)**
(RMSE)	XGB	23.11 (0.26)	20.43 (0.34)	14.39 (0.2)	14.28 (0.23)	**14.05 (0.29)**
PortSecStud	GLM	3.86 (0.17)	3.76 (0.11)	1.57 (0.1)	1.58 (0.1)	**1.56 (0.1)**
(RMSE)	XGB	3.86 (0.17)	3.83 (0.15)	**1.49 (0.09)**	**1.54 (0.14)**	**1.54 (0.05)**
xAPI-Edu	GLM	0.2 (0.02)	0.39 (0.04)	**0.74 (0.03)**	**0.74 (0.05)**	**0.74 (0.06)**
(F1)	XGB	0.2 (0.02)	0.54 (0.03)	0.74 (0.03)	**0.75 (0.05)**	**0.78 (0.05)**
OULAD	GLM	0.81 (0.0)	0.81 (0.0)	**0.87 (0.01)**	**0.87 (0.01)**	**0.87 (0.01)**
(F1)	XGB	0.81 (0.0)	0.81 (0.0)	**0.91 (0.0)**	**0.91 (0.0)**	**0.91 (0.0)**

Table 3 shows the results for different data subsets as defined in Subsection 5.1. We report the mean of the performance metric as well as the standard deviations. Bold numbers represent

[2] To view the distribution of values of the demographic features, see Appendix B.

statistically significant best numbers per row, indicating the best feature subsets for the dataset and method. The columns indicate what feature subsets were used, with “All” meaning that all feature subsets, including the “other” features, were used.

6.1.1. Predictive Capability of Demographic Features

For OULAD, no difference can be seen between the baseline (i.e., simply predicting the mode) and using solely demographic features for prediction. Hence, predicting that every student passes the course works equally well as training a model only using demographic features. For the PortSecStud dataset, the improvement over the baseline (i.e., simply predicting the mean) is small, such that the usefulness of the demographic features can also be considered small for this dataset. For the Engineering dataset, there is a considerable improvement over the baseline, and for the xAPI-Edu-Data, the improvement over the baseline is the largest. Hence, for these two datasets, it can be said that demographic features impact performance. Considering **RQ1**, we conclude that demographic characteristics can be used to explain differences in student achievement. However, this conclusion does not hold in every setting and for every type of demographic characteristic. Regarding fairness, it seems to be the case that demographic features are not always but sometimes correlated with the target, meaning that the notion of fairness through unawareness is interesting for the setting of at-risk prediction but also shows that we perhaps should not necessarily aim for highly accurate predictions at all given the biased target.

6.1.2. The Role of Study-Related Features

For all datasets, using only study-related features achieves far better performance than using only demographic features. Using only study-related features achieves approximately the same performance as additionally considering demographic features in almost every setting. Only for XGBoost on xAPI-Edu-Data and GLM on PortSecStud is there a noteworthy mean difference, although it is small in absolute values. The results for PortSecStud show that using all information leads to the best performance. This result indicates that the “other” features are important. As these might be closely related to demographic features or might even be counted as demographic features using a different definition, we might be inclined to say that demographic features are important even if we have study-related features. This conclusion that demographic information perhaps matters is only true for GLM, though. In general, these results confirm the hypothesis that study-related information contains information about demographic characteristics and that this information might be used to predict the target. What follows is that as soon as meaningful information about the student’s activity and/or previous performance is available, the demographic features are no longer required for accurate predictions. Hence, considering **RQ2**, demographic characteristics are generally not helpful for predictive accuracy anymore if study-related information is available. From a fairness perspective, this means that the resulting model may still be biased even if we remove demographic features.

6.2. Feature Importance of Demographic Data

The previous subsections have provided clear evidence that demographic features are not necessary for at-risk predictions when sufficient information about students’ study activities or previous performance is available from a performance point of view. However, our theoretical considerations indicate that demographic features are likely correlated with other study-related features and the target. Thus, it is possible that models use these demographic features when

Table 4: Means and standard deviations of relative feature importances of demographic data compared to the rest of the data in the model on different data subsets over all folds.

Dataset		Demo only	Demo + Study	All
Engineering	GLM	1.0 (0.0)	0.25 (0.03)	0.23 (0.02)
	XGB	1.0 (0.0)	0.23 (0.06)	0.19 (0.09)
PortSecStud	GLM	1.0 (0.0)	0.08 (0.04)	0.03 (0.02)
	XGB	1.0 (0.0)	0.26 (0.04)	0.16 (0.04)
xAPI-Edu	GLM	1.0 (0.0)	0.32 (0.14)	0.23 (0.1)
	XGB	1.0 (0.0)	0.26 (0.03)	0.21 (0.02)
OULAD	GLM	1.0 (0.0)	0.13 (0.0)	0.13 (0.0)
	XGB	1.0 (0.0)	0.08 (0.0)	0.08 (0.01)

Table 5: Means and standard deviations of the effect of demographic features on the predictions over all folds. For regression datasets (Engineering, PortSecStud), the mean absolute difference between predictions with and without demographic data is reported. For classification datasets (xAPI-Edu, OULAD), the percentage of predictions that change when excluding the demographic features from the model is reported.

Dataset	Effect of demographics
Engineering	0.97 (0.07)
PortSecStud	0.28 (0.06)
xAPI-Edu	0.11 (0.05)
OULAD	0.33 (0.01)

they are included in the training data. To further inspect whether this is the case, we analyze the linear models' learned coefficients and the XGBoost models' feature importances as described in Subsection 5.6. Table 4 states the means and standard deviations of relative feature importance of demographic data compared to the rest of the data. These results allow us to assess how large their influence on the prediction is. The results show that despite the fact that an equally good model could have usually been learned for all models without demographic features, those are still used for all models and datasets. Even for the PortSecStud dataset and the OULAD dataset, where we previously found that demographic features do not help at all compared to the naive baseline, the features are still used. For the XGBoost model trained on study-related and demographic data of the PortSecStud dataset, the demographic information even accounts for 26% of the feature importance despite not being necessary to achieve the performance. This result might be an indicator of the confounder relationship we theorized about before. The confounder relationship would mean that while the target itself is not influenced by demographic variables, other predictive features that are also correlated with the target are, and the models correct for this influence by also using demographic features. For the other datasets, where demographic data was useful, it is likely the case that demographic data is used because it can provide information more directly than through other predictive features. Furthermore, Table 5 shows that in every case, the utilization of demographic data directly affects the predictions of the models. The table shows the mean absolute differences between predictions when using and not using demographic features for the regression tasks and the percentage of instances that receive a different prediction for the classification tasks. For regression, the effect is not large, considering the scales of the targets. Nevertheless, it might be important for some students or some student subgroups that are too small to account for a large mean difference. For classification, the impact of demographic features on actual predictions is large.

Our answer to **RQ3** is that just using all features leads to models that use information from demographic features. That, combined with the previous results of this section, shows that the notion of fairness through unawareness is generally important to consider and that removing demographic features makes a noticeable difference.

6.3. Bias Indication Through Fairness Metrics

Table 6: The table reports the means and standard deviations of the demographic parity ratio of all demographic features and then of 5CV fairness metrics results on different data subsets. The values are rounded to the second decimal place. The smallest DP for each row is bold.

Dataset		Demo only	Study Only	Demo+Study	All
Engineering	GLM	0.92 (0.02)	**0.91 (0.00)**	**0.91 (0.01)**	**0.91 (0.01)**
	XGB	**0.89 (0.02)**	0.90 (0.00)	0.90 (0.01)	0.90 (0.00)
PortSecStud	GLM	0.88 (0.05)	**0.86(0.02)**	**0.86 (0.04)**	**0.86(0.04)**
	XGB	**0.86 (0.02)**	0.87 (0.02)	0.87 (0.03)	0.87 (0.01)
xAPI-Edu	GLM	**0.16 (0.04)**	0.54 (0.04)	0.33 (0.02)	0.40 (0.03)
	XGB	**0.27 (0.05)**	0.59 (0.05)	0.42 (0.04)	0.52 (0.05)
OULAD	GLM	0.79 (0.01)	0.80 (0.00)	**0.78 (0.00)**	**0.78 (0.00)**
	XGB	**0.76 (0.01)**	0.82 (0.01)	0.79 (0.00)	0.79 (0.00)

Table 7: Feature and value pairs selected according to the following mechanism: Feature for which the largest change occurred between Demo+Study and Study only – the Demo+Study value and difference (+ meaning that the value increased towards Study only, - the opposite) is reported in this order; feature with smallest and largest demographic parity ratio. The last column shows the number of demographic features with a ratio below 0.8 for Study-only. All values were averaged across 5CV and rounded to the second decimal place.

Dataset		Largest Difference	Smallest Value	Largest Value	N < 0.8
Engineering	GLM	SCHOOL_NAME: 0.56 / +0.02	SCHOOL_NAME: 0.58	GENDER: 0.99	1
	XGB	SCHOOL_NAME: 0.43 / +0.05	SCHOOL_NAME: 0.48	GENDER: 0.99	2
PortSecStud	GLM	traveltime: 0.84 / -0.02	age: 0.69	paid: 0.97	3
	XGB	traveltime: 0.87 / -0.07	age: 0.74	paid: 0.97	4
xAPI-Edu	GLM	Relation: 0.17 / +0.42	PlaceofBirth: 0.39	gender: 0.71	4
	XGB	Relation: 0.30 / +0.33	PlaceofBirth: 0.33	gender: 0.85	3
OULAD	GLM	highest_education: 0.56 / +0.07	highest_education: 0.63	gender: 0.97	3
	XGB	highest_education: 0.61 / +0.04	highest_education: 0.65	gender: 0.97	4

Table 6 shows the average demographic parity ratio for each model and data subset across all demographic features. By looking at the table, we can observe two major aspects: First, most models apart from those concerned with xAPI-Edu look rather unbiased when looking at the demographic parity ratio – even those just using demographic features. Second, including or not including demographic features barely makes a difference except for xAPI-Edu. The first point can very simply be explained by the fact that most demographic features do not matter for the predictions at all, and, therefore, there is barely a bias coming from them increasing the average.

There are, of course, other demographic features that indicate a bias for every model, as can be seen in Table 7. The observation that models trained on xAPI-Edu indicate bias much more strongly might be because the dataset contains only a few demographic features, but those might almost all be very relevant for the prediction. Our second observation that removing demographic features barely makes a difference except for xAPI-Edu might, in part, still be because most features do not matter much at all. However, going back to our theoretical considerations in Section 2.3, this result also corresponds to our idea that the effect of the demographic features is still encoded in the other predictive features.

For Table 7, the last column shows the number of demographic features for the Study-only subset that have a ratio below 0.8, indicating a bias. We can see that always some, but usually not all, features indicate a bias. The second to last column shows the feature with the highest demographic parity ratio, i.e., with the least amount of bias. Here, we can see that, interestingly, gender is the least biased feature for three datasets, for both XGBoost and GLM. The middle left column shows the smallest value for each dataset and indicates a bias in all cases (i.e., a value smaller than 0.8). Interestingly, the variables with the smallest and largest values, respectively, for GLM and XGBoost, always correspond, e.g., for PortSecStud, “age”” has the smallest value and “paid” the largest for both GLM and XGBoost.

The first column shows the largest difference between the demographic parity ratio of the Study-only and the Study+Demo subset. We can observe that for two datasets, Engineering and OULAD, the smallest value already experienced the largest increase, suggesting that those features indicating a bias benefit the most from removing the demographic features. It should also be noted, however, that the change is not sizable enough to lift the values above the 0.8 threshold for any of the datasets. In general, the differences are not very large – with the exception of xAPI-Edu – which, again, points to the idea that the information of the demographic features relevant for the prediction is encoded in the other predictive features. Related to this conclusion, xAPI-Edu might experience a stronger increase because – due to the small sample size – the models cannot properly learn the more complicated functions to reproduce the biases coming from the effect of demographic features on the other predictive features. A final observation is that the largest change for PortSecStud is in the opposite direction: the value for traveltime increases when removing the demographic features indicating a larger bias for both models. This result fits what we already theorized about, namely that for this dataset, the target is unaffected by at least some demographic features, but other predictive features are influenced by these demographic features and are correlated with the target through a confounder. When including these demographic features, they correct their own influence. But when we exclude them, the prediction is more biased. For other demographic features in this dataset, this effect probably does not hold, considering that the overall change is not significant.

All in all, **RQ4** is answered with no; there is no substantive difference with regard to demographic parity when it comes to removing demographic features. We can see that while most demographic features do not lead to a bias indication at all, some do. For those features, removing the demographic feature usually but not always has a positive impact, but not such a large one that we would answer **RQ4** in the affirmative. This conclusion is in line with our idea that other predictive features encode the information of relevant demographic features, meaning that removing the demographic features has only a small effect. Notably, for PortSecStud, it might even be the case that removing some of the demographic features leads to a more biased model.

7. Discussion

Our evaluation shows that using demographic features generally does not lead to better model performance as long as we include study-related features. If we leave demographic features in, however, they will be used by the models. Hence, we generally advise in favor of removing demographic features (as advised by the notion of “fairness through unawareness”). However, doing so does not guarantee unbiased and fair models (opposing the view of “fairness through unawareness”). Our results clearly show that at least some demographic features still have an influence on the target when demographic features are removed. This finding conforms to our theory presented in Subsection 2.3 that indirect, mediating effects are at play. Moreover, there may be some features for which their removal means that the model becomes more biased in their regard. This is because the features may not influence the target directly, but do influence other features correlated with the target. The features correct their own influence when used in models, but of course cannot do this if they are not included. It seems to be the case that this happens rarely. Still, it should be considered and shows that the topic of removing demographic features for predictive models is difficult and that fairness through unawareness does not necessarily match other notions of fairness as described in Subsection 2.2. In general, we still do advocate the removal of demographic features to prevent them from being used for predictions, which is usually not considered fair. We will now turn to the implications our results have for practitioners and researchers.

7.1. The Importance of Demographic Features for Prediction

Clearly, demographic features only sometimes make models perform better in comparison to only having study-related information available. Yet, models use them if they are included. What follows are two notable consequences for practitioners and researchers.

First, often, the target itself is also biased. Going back to the example presented in Section 2.3, if the target feature is admittance and men are perceived as more able, resulting in more men being admitted, then the target contains a bias. Hence, we do not necessarily want to achieve models with the best possible performance, as that would include bias. Usually, we chase for better performance, but our paper indicates that this is not necessarily the best option. We recommend that practitioners and researchers think critically about what they want their model to be and whether it is possible that by simply trying to be as accurate as possible, biases are reinforced.

Second, practitioners and researchers should focus on gathering and using study-related data. These features are clearly very important, and while demographic features might also influence these features, their usage is at least comparatively fairer, and existing biases in them can be mitigated using bias mitigation strategies (Deho et al., 2022). Demographic features, on the other hand, can be left out from a performance perspective. Note that in the cases for which leaving them out leads to a decrease in performance, the true target is likely biased, reinforcing the first point.

Leaving them out of predictive models, however, does not mean that we should not consider them at all in the context of predictive models. As a matter of fact, it is important to collect demographic information still and to use it to evaluate the models regarding different fairness notions as we did in this study (Andrus et al., 2021).

7.2. The Importance of Demographic Features for Non-Prediction Models

We want to highlight again that the recommendation not to include demographic features does not mean that we should never explore the impact of them on academic achievement or that demographic features are not important. On the contrary, it is very important to investigate how demographic characteristics impact academic achievement so that we can intercept mechanisms that would lead to disadvantages for certain populations (Paquette et al., 2020). We also want to highlight that for all datasets, only a few demographic features were relevant enough for the predictions that they produced a bias. Learning more about which features matter in which settings and how they matter is an interesting topic worth researching.

Hence, we certainly do not want to discourage research on causal mechanisms of demographic characteristics' impact on academic achievement. Rather, we want to highlight the importance of thinking about demographic features. The fact that at least some other predictive features transport information about demographic features, even if we do not include demographic features, shows the necessity to think about how demographic features are embedded in the causal mechanisms generating the data.

7.3. The Importance of Demographic Features for Fairness

For fairness, our implications are less clear-cut and cannot be general. We may say that our results do not support the claim in Yu et al. (2021) that sensitive features should be included for fairness reasons if the prediction accuracy itself is equal. Table 5 shows that using demographic features changed the prediction for some students; this is an indication that the models using demographic features are, indeed, unfair (Mehrabi et al., 2021). Therefore, leaving them in would probably not result in an equally fair model when considering the notion of fairness through unawareness as explained in Subsection 2.2.

However, this result does not mean that models are fair according to all notions of fairness when removing demographic features. As we also showed in our results, models are still often biased with regard to at least one demographic feature due to the other features being correlated with the demographic feature, making models biased according to DP. Depending on the application and setting, it might be necessary to take further steps, such as employing bias mitigation strategies (Deho et al., 2022; Bird et al., 2021). At the very least, this possibility is something that researchers and, in particular, practitioners should critically think about.

To make matters even more complicated, removing the demographic features may, in certain settings, even make the model more biased. This surprising effect clearly highlights the need to think about the causal mechanisms underlying our data. Regarding the causal mechanisms, we want to point practitioners to a recent article by Cohausz et al. (2024), which depicts different causal relationships similar to the ones discussed in Subsection 2.3 and how they affect algorithmic bias as well as our perception of fairness, which may help researchers and practitioners with thinking about fairness and deciding on courses of action on a case-by-case basis. What we ultimately judge as fair or unfair is both context-specific and subjective.

7.4. Limitations of Our Study and Future Work

Despite our solid results, it is important to note certain limitations. We only used four datasets and two types of models to test our hypotheses. As these datasets are diverse (online, offline, different countries, different levels of education), as are the model types (linear, non-linear), we believe that our main findings are still very reliable. Nonetheless, future work should investigate whether the findings hold when using other datasets and models.

Additionally, we did not investigate whether the models' feature importance may change when using different encoding methods. This may be the case when the encoding methods learn that demographic features are not necessary for the prediction. However, given the correlation between demographic features and the target, it is unlikely that different encodings would lead to models not contributing importance to demographic features at all. Still, this should also be investigated in the future.

Because it is not the major focus of our study, we have not investigated the relationship between the importance of activity- and performance-related features. Future research could investigate what is more important and how the two feature subsets relate to each other.

Furthermore, as already stated in the introduction, this study only concerned itself with one task of EDM, but other problems handled in EDM research (e.g., course recommendations) also make use of demographic features. We argue that similar studies should be conducted for other tasks in EDM as well and hope that our paper can be an orientation to these future research endeavors.

Finally, and as already mentioned, we think that our study highlights the need to better understand the impact of demographic features on performance and activity-related features. Doing so could also help reduce the impact of demographic features on other predictive features. Therefore, we highly encourage research that concerns itself with modeling the causal relationships among those features and investigating related bias mitigation strategies.

8. Conclusion

In this paper, we theoretically reasoned how demographic features may impact the predictions of ML models regarding both performance and algorithmic bias. Our subsequent empirical evaluation shows strong evidence that demographic features do not typically increase a model's performance on at-risk prediction as long as study-related information is available. However, demographic features are still used by the models when available. Because of this, we advise leaving out demographic features. Nonetheless, models may still be biased because non-sensitive predictive features are also correlated with the demographic features transporting their information. This highlights the need to think critically about fairness concerns and take further measures if necessary. In particular, our paper shows that investigating the causal mechanisms of how demographic features impact academic achievement is worthwhile and should be encouraged, with the aim of achieving fairer models.

References

AL-ZAWQARI, A. AND VANDERSTEEN, G. 2022. Investigating the role of demographics in predicting high achieving students. In *23rd International Conference on Artificial Intelligence in Education*, M. M. Rodrigo, N. Matsuda, A. I. Cristea, and V. Dimitrova, Eds. Springer, Durham, UK, 440–443.

ALTURKI, S., HULPUȘ, I., AND STUCKENSCHMIDT, H. 2022. Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning 27,* 1, 275–307.

AMRIEH, E. A., HAMTINI, T., AND ALJARAH, I. 2016. Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application 9,* 8, 119–136.

ANDRUS, M., SPITZER, E., BROWN, J., AND XIANG, A. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. Association for Computing Machinery, New York, USA, 249–260.

AZHAR, M., NADEEM, S., NAZ, F., PERVEEN, F., AND SAMEEN, A. 2014. Impact of parental education and socio-economic status on academic achievements of university students. *European Journal of Psychological Research 1,* 1, 1–9.

BAKER, R. S., ESBENSHADE, L., VITALE, J., KARUMBAIAH, S., ET AL. 2023. Using demographic data as predictor variables: a questionable choice. *Journal of Educational Data Mining 15,* 2, 22–52.

BATOOL, S., RASHID, J., NISAR, M. W., KIM, J., MAHMOOD, T., AND HUSSAIN, A. 2021. A random forest students' performance prediction (rfspp) model based on students' demographic features. In *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)*. IEEE, Karachi, Pakistan, 1–4.

Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D. D. 2015. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery 8,* 1, 014008.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. 2021. Fairlearn: A toolkit for assessing and improving fairness in ai. Tech. rep., Microsoft, Tech. Rep. MSR-TR-2020-32.

Bloodhart, B., Balgopal, M. M., Casper, A. M. A., Sample McMeeking, L. B., and Fischer, E. V. 2020. Outperforming yet undervalued: Undergraduate women in stem. *Plos one 15,* 6, e0234685.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports 12,* 1, 4209.

Caton, S. and Haas, C. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys 56,* 7, 1–38.

Cohausz, L., Kappenberger, J., and Stuckenschmidt, H. 2024. What fairness metrics can really tell you: A case study in the educational domain. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. Association for Computing Machinery, New York, USA, 792–799.

Cohausz, L., Tschalzev, A., Bartelt, C., and Stuckenschmidt, H. 2023. Investigating the importance of demographic features for edm-predictions. In *16th International Conference on Educational Data Mining*. International Educational Data Mining Society, Bengaluru, India.

Cortez, P. and Silva, A. M. G. 2008. Using data mining to predict secondary school student performance. Tech. rep., EUROSIS-ETI.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., and Alowibdi, J. S. 2017. Predicting student performance using advanced learning analytics. In *Proceedings of the 26th international conference on world wide web companion*. nternational World Wide Web Conferences Steering Committee, Republic and Canton of GenevaSwitzerland, 415–421.

Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L., and Duy Le, T. 2022. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology 53,* 4, 822–843.

Delahoz-Dominguez, E., Zuluaga, R., and Fontalvo-Herrera, T. 2020. Dataset of academic performance evolution for engineering students. *Data in brief 30*, 105537.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. Association for Computing Machinery, New York, USA, 214–226.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems 35*, 507–520.

HILL, C., CORBETT, C., AND ST ROSE, A. 2010. *Why so few? Women in science, technology, engineering, and mathematics.* ERIC, 1111 Sixteenth Street NW, Washington, DC 20036.

HOFFAIT, A.-S. AND SCHYNS, M. 2017. Early detection of university students with potential difficulties. *Decision Support Systems 101*, 1–11.

HU, Q. AND RANGWALA, H. 2020. Towards fair educational data mining: A case study on detecting at-risk students. In *13th International Conference on Educational Data Mining*. International Educational Data Mining Society, Ifrane, Morocco.

JHA, N. I., GHERGULESCU, I., AND MOLDOVAN, A.-N. 2019. Oulad mooc dropout and result prediction using ensemble, deep learning and regression techniques. In *Proceedings of the 11th International Conference on Computer Supported Education*. Springer Nature, Heraklion, Crete, Greece, 154–164.

KHASANAH, A. U. ET AL. 2017. A comparative study to predict student's performance using educational data mining techniques. In *IOP Conference Series: Materials Science and Engineering*. Vol. 215. IOP Publishing, Bristol, UK, 012036.

KIZILCEC, R. F. AND LEE, H. 2022. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*. Routledge, New York, USA, 174–202.

KUZILEK, J., HLOSTA, M., AND ZDRAHAL, Z. 2017. Open university learning analytics dataset. *Scientific data 4,* 1, 1–8.

MEHRABI, N., MORSTATTER, F., SAXENA, N., LERMAN, K., AND GALSTYAN, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54,* 6, 1–35.

MICCI-BARRECA, D. 2001. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter 3,* 1, 27–32.

MIGUÉIS, V. L., FREITAS, A., GARCIA, P. J., AND SILVA, A. 2018. Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems 115*, 36–51.

PAQUETTE, L., OCUMPAUGH, J., LI, Z., ANDRES, A., AND BAKER, R. 2020. Who's learning? using demographics in edm research. *Journal of Educational Data Mining 12,* 3, 1–30.

PARGENT, F., PFISTERER, F., THOMAS, J., AND BISCHL, B. 2022. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics 37,* 5, 1–22.

PEARL, J. 2009. *Causality*. Cambridge university press, Cambridge, UK.

PROKHORENKOVA, L., GUSEV, G., VOROBEV, A., DOROGUSH, A. V., AND GULIN, A. 2018. Catboost: unbiased boosting with categorical features. In *Advances in neural information processing systems*. IEEE, Montreal, Canada.

SALKIND, N. J. 2010. *Encyclopedia of research design*. Vol. 1. sage, Thousand Oaks, California, USA.

SHWARTZ-ZIV, R. AND ARMON, A. 2022. Tabular data: Deep learning is not all you need. *Information Fusion 81*, 84–90.

SWEENEY, M., LESTER, J., RANGWALA, H., JOHRI, A., ET AL. 2016. Next-term student performance prediction: A recommender systems approach. *Journal of Educational Data Mining 8,* 1, 22–51.

TOMASEVIC, N., GVOZDENOVIC, N., AND VRANES, S. 2020. An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & education 143*, 103676.

TRSTENJAK, B. AND DONKO, D. 2014. Determining the impact of demographic features in predicting student success in croatia. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, Opatija, Croatia, 1222–1227.

WU, J., CHEN, X.-Y., ZHANG, H., XIONG, L.-D., LEI, H., AND DENG, S.-H. 2019. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology 17,* 1, 26–40.

YU, R., LEE, H., AND KIZILCEC, R. F. 2021. Should college dropout prediction models include protected attributes? In *Proceedings of the eighth ACM conference on learning@ Scale*. Association for Computing Machinery, New York, USA, 91–100.

ZHAO, Y., XU, Q., CHEN, M., AND WEISS, G. 2020. Predicting student performance in a master's program in data science using admissions data. In *13th International Conference on Educational Data Mining*. International Educational Data Mining Society, Ifrane, Morocco.

A. Appendix A

The results for the different encoding techniques can be found in Table 8. Clearly, all encoding techniques perform similarly.

Table 8: Means and standard deviations of 5CV Performance results. Mean squared error is reported for Engineering and PortSecStud and F1-score for xAPI-Edu and OULAD. Results per row for encodings that are not significantly different from the best encoding in a paired t-test (alpha=0.05) are highlighted in bold.

Dataset		Baseline	Ignore	OHE	Target	Ordinal	Catboost	5CV-GLMM
Engineering	GLM	23.11 (0.26)	14.36 (0.26)	**14.11 (0.29)**	14.55 (0.23)	14.36 (0.25)	14.22 (0.29)	**14.13 (0.29)**
	XGB	23.11 (0.26)	14.28 (0.25)	14.11 (0.26)	14.38 (0.23)	14.15 (0.25)	14.17 (0.28)	**14.04 (0.28)**
PortSecStud	GLM	3.86 (0.17)	**1.55 (0.1)**	**1.55 (0.1)**	**1.55 (0.1)**	**1.55 (0.1)**	**1.55 (0.1)**	**1.55 (0.1)**
	XGB	3.86 (0.17)	**1.51 (0.12)**	1.51 (0.08)	**1.46 (0.06)**	**1.52 (0.06)**	**1.57 (0.08)**	**1.5 (0.06)**
xAPI-Edu	GLM	0.2 (0.02)	**0.75 (0.06)**	**0.76 (0.03)**	**0.75 (0.06)**	**0.73 (0.05)**	**0.73 (0.09)**	**0.75 (0.06)**
	XGB	0.2 (0.02)	**0.76 (0.05)**	**0.78 (0.04)**	**0.78 (0.06)**	**0.78 (0.08)**	0.72 (0.07)	**0.76 (0.02)**
OULAD	GLM	0.81 (0.0)	**0.87 (0.01)**	**0.87 (0.01)**	**0.87 (0.01)**	**0.87 (0.01)**	**0.87 (0.01)**	**0.87 (0.01)**
	XGB	0.81 (0.0)	**0.91 (0.0)**	**0.91 (0.0)**	**0.91 (0.0)**	**0.91 (0.0)**	**0.91 (0.0)**	**0.91 (0.0)**

B. Appendix B

Appendix B shows the distribution of the values for the demographic features in all datasets.

Table 9: Distribution of the values of the demographic features of Engineering in percent.

Feature	Value	Percentage
Gender	Male	0.59
	Female	0.41
Education Father	Complete professional education	0.24
	Complete Secundary	0.23
	Complete technique or technology	0.10
	Incomplete Secundary	0.09
	Postgraduate education	0.09
	Complete primary	0.07
	Incomplete primary	0.06
	Incomplete Professional Education	0.04
	Not sure	0.04
	Other	0.04
	Incomplete technical or technological	0.03
Education Mother	Complete Secundary	0.25
	Complete professional education	0.25
	Complete technique or technology	0.12
	Incomplete Secundary	0.09
	Postgraduate education	0.08
	Complete primary	0.06
	Incomplete primary	0.04
	Incomplete Professional Education	0.04
	Other	0.03
	Incomplete technical or technological	0.03
	Not sure	0.01
Occupation Father	Independent	0.23
	Technical or professional level employee	0.15
	Operator	0.12
	Other occupation	0.09
	Executive	0.09
	Other	0.08
	Independent professional	0.07
	Small entrepreneur	0.06
	Retired	0.04
	Entrepreneur	0.04
	Auxiliary or Administrative	0.03
	Home	0.01
	Home	0.38
	Technical or professional level employee	0.14

Occupation Mother	Independent	0.09
	Auxiliary or Administrative	0.07
	Executive	0.06
	Independent professional	0.06
	Operator	0.06
	Other occupation	0.05
	Small entrepreneur	0.04
	Other	0.03
	Entrepreneur	0.02
	Retired	0.01
Stratum	Stratum 3	0.33
	Stratum 2	0.32
	Stratum 1	0.14
	Stratum 4	0.13
	Stratum 5	0.05
	Stratum 6	0.03
SISBEN	Unclassified	0.61
	Level 2	0.17
	Level 1	0.17
	Level 3	0.05
	Other	0.01
People in House	Four	0.38
	Five	0.23
	Three	0.19
	Six	0.09
	Two	0.05
	Seven	0.03
	Eight	0.01
	Nine	0.01
	Ten	<0.01
	Twelve or more	<0.01
	One	<0.01
Internet	Yes	0.79
TV	Yes	0.85
Computer	Yes	0.82
Washing Machine	Yes	0.62
Oven	Yes	0.69
Car	Yes	0.53
DVD	Yes	0.75
Fresh Water	Yes	0.97
Phone	Yes	0.96
Mobile	Yes	0.71
	Between 1 and less than 2 LMMW	0.31
	Between 2 and less than 3 LMMW	0.22
	Between 3 and less than 5 LMMW	0.18

Revenue	less than 1 LMMW	0.08
	Between 5 and less than 7 LMMW	0.08
	10 or more LMMW	0.06
	Between 7 and less than 10 LMMW	0.04
	Other	0.02
Job	No	0.96
	< 20h/week	0.02
	> 20h/week	0.01
	Other	0.01
School	Private	0.53
	Public	0.47

Table 10: Distribution of the values of the demographic features of PortSecStud in percent.

Feature	Value	Percentage
School	GP	0.74
	MS	0.26
gender	Female	0.57
	Male	0.43
Age	16	0.27
	17	0.27
	18	0.21
	15	0.19
	19	0.05
	20	0.01
	21	< 0.01
	22	< 0.01
Address	Urban	0.73
	Rural	0.27
Family Size	> 3	0.71
	$<= 3$	0.29
Mother Education	4	0.29
	2	0.28
	3	0.23
	1	0.19
	0	0.01
Father Education	2	0.31
	1	0.25
	3	0.22
	4	0.21
	0	0.01
Mother Job	Other	0.38
	Services	0.23
	At home	0.19
	Teacher	0.12
	Health	0.08
Father Job	Other	0.56
	Services	0.28
	teacher	0.07
	At home	0.06
	Health	0.04
Guardian	Mother	0.70
	Father	0.23
	Other	0.07
Travel Time	1	0.60
	2	0.31

	3	0.07
	4	0.02
School Money Support	No	0.89
Family Money Support	No	0.61
Paid Classes	Yes	0.79
Internet	Yes	0.79
Family Relationship	4	0.49
	5	0.27
	3	0.16
	2	0.05
	1	0.03
Health	5	0.38
	3	0.21
	4	0.17
	1	0.13
	2	0.12

Table 11: Distribution of the values of the demographic features of xAPI-Edu in percent.

Feature	Value	Percentage
Gender	Male	0.64
	Female	0.36
Nationality	Kuwait	0.38
	Jordan	0.36
	Palestine	0.06
	Iraq	0.05
	Lebanon	0.05
	Tunisia	0.03
	Saudi-Arabia	0.02
	Egypt	0.02
	Syria	0.01
	USA	0.01
	Iran	0.01
	Lybia	0.01
	Morocco	0.01
	Venezuela	< 0.01
Place of Birth	Kuwait	0.38
	Jordan	0.37
	Iraq	0.05
	Lebanon	0.04
	USA	0.03
	Saudi-Arabia	0.03
	Palestine	0.02
	Egypt	0.02
	Tunisia	0.02
	Syria	0.01
	Iran	0.01
	Lybia	0.01
	Morocco	0.01
	Venezuela	< 0.01
Parent Responsible	Father	0.59
	Mother	0.41

Table 12: Distribution of the values of the demographic features of OULAD in percent.

Feature	Value	Percentage
Gender	Male	0.55
	Female	0.45
Region	Scotland	0.12
	East Anglian Region	0.10
	London Region	0.10
	South Region	0.09
	North Western Region	0.09
	West Midlands Region	0.08
	South West Region	0.07
	East Midlands Region	0.07
	South East Region	0.06
	Wales	0.06
	Yorkshire Region	0.06
	North Region	0.06
	Ireland	0.04
Highest Education	A Level or Equivalent	0.43
	Lower Than A-Level	0.40
	HE Qualification	0.15
	No Formal	0.01
	Post Graduate Qualification	0.01
IMD-Band	20-30	0.11
	30-40	0.11
	10-20	0.11
	0-10	0.10
	40-50	0.10
	50-60	0.10
	60-70	0.09
	70-80	0.09
	80-90	0.08
	90-100	0.08
Age Band	0-35	0.70
	35-55	0.29
	>55	0.01
Disability	No	0.90