education sciences

MDPI

Article

Assessing Disparities in Predictive Modeling Outcomes for College Student Success: The Impact of Imputation Techniques on Model Performance and Fairness

Nazanin Nezami [1], Parian Haghighat [1], Denisa Gándara [2] and Hadis Anahideh [1,*]

[1] Mechanical and Industrial Engineering Department, University of Illinois Chicago, Chicago, IL 60612, USA; nnezam2@uic.edu (N.N.); phaghi3@uic.edu (P.H.)
[2] Department of Educational Leadership and Policy, University of Texas at Austin, Austin, TX 78712, USA; denisa.gandara@austin.utexas.edu
* Correspondence: hadis@uic.edu

Abstract: The education sector has been quick to recognize the power of predictive analytics to enhance student success rates. However, there are challenges to widespread adoption, including the lack of accessibility and the potential perpetuation of inequalities. These challenges present in different stages of modeling, including data preparation, model development, and evaluation. These steps can introduce additional bias to the system if not appropriately performed. Substantial incompleteness in responses is a common problem in nationally representative education data at a large scale. This can lead to missing data and can potentially impact the representativeness and accuracy of the results. While many education-related studies address the challenges of missing data, little is known about the impact of handling missing values on the fairness of predictive outcomes in practice. In this paper, we aim to assess the disparities in predictive modeling outcomes for college student success and investigate the impact of imputation techniques on model performance and fairness using various notions. We conduct a prospective evaluation to provide a less biased estimation of future performance and fairness than an evaluation of historical data. Our comprehensive analysis of a real large-scale education dataset reveals key insights on modeling disparities and the impact of imputation techniques on the fairness of the predictive outcome under different testing scenarios. Our results indicate that imputation introduces bias if the testing set follows the historical distribution. However, if the injustice in society is addressed and, consequently, the upcoming batch of observations is equalized, the model would be less biased.

Keywords: machine learning; fairness of predictive outcomes; imputation; prospective evaluation; education analytics

check for updates

Citation: Nezami, N.; Haghighat, P.; Gándara, D.; Anahideh, H. Assessing Disparities in Predictive Modeling Outcomes for College Student Success: The Impact of Imputation Techniques on Model Performance and Fairness. *Educ. Sci.* **2024**, *14*, 136. https://doi.org/10.3390/educsci14020136

Received: 2 October 2023
Revised: 6 January 2024
Accepted: 11 January 2024
Published: 29 January 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Predictive analytics has become an increasingly hot topic in higher education. Predictive analytics tools have been used to predict various measures of student success (e.g., course completion, retention, and degree attainment) by mapping the input set of attributes of individuals (e.g., the student's high school GPA and demographic features) with their outcomes (e.g., college credits accumulated) [1]. Campus officials have used these predictions to guide decisions surrounding college admissions and student-support interventions, such as providing more intensive advising to certain students [1].

Despite the potential for predictive analytics, there is a critical disconnection between predictive analytics in higher education research and their accessibility in practice. There are two major barriers to existing uses of predictive analytics in higher education that cause this disconnection are formidable challenges to the widespread adoption of predictive analytics in higher education. The first is the lack of democratization in its deployment, which can limit its accessibility among practitioners and education researchers. The second

is the potential to perpetuate existing inequalities, which could exacerbate disparities in outcomes based on race, gender, or socioeconomic status. Consequently, predictive analytics in higher education must be carefully implemented to avoid further widening the gap between the haves and have-nots.

First, education researchers and policymakers face many challenges in deploying predictive and statistical techniques in practice. These challenges present in different steps of modeling including data cleaning (e.g., imputation), identifying the most important attributes associated with success, selecting the correct predictive modeling technique, and calibrating the hyperparameters of the selected model. Each of these steps can introduce additional bias to the system if not appropriately performed [2]. Missing values are the frequent latent causes behind many data analysis challenges. Most large-scale and nationally representative education datasets suffer from a significant number of incomplete responses from the research participants. While many education-related studies address the challenges of missing data [3–5], little is known about the impact of handling missing values on the fairness of predictive outcomes in practice. To date, few works have studied the impact of data preparation on the unfairness of the predictive outcome in a limited setting [6] or use merely a single notion of fairness metrics [7].

Second, predictive models rely on historical data and have the potential to exacerbate social inequalities [1,8]. Over the last decade, researchers have realized that disregarding the consequences and especially the societal impact of algorithmic decision making might negatively impact individuals' lives, especially those who are most marginalized in society. COMPAS, a criminal justice support tool, was found to be decidedly biased against Black people [9]. Colleges and universities have been using risk algorithms to evaluate their students. Recently, Markup investigated four major public universities and found that EAB's Navigate software was racially biased [10]. Ensuring fair and unbiased assessment, however, is complex and it requires education researchers and practitioners to undergo a comprehensive algorithm audit to ensure the technical correctness and social accountability of their algorithms.

It is imperative that predictive models are designed with careful attention to their potential social consequences. A wave of fair decision-making algorithms and, in particular, fair machine learning (ML) models for prediction have been proposed in recent years [11,12]. Nevertheless, most of the proposed research either deals with inequality in the pre-processing or post-processing steps or considers a model-based in-processing approach. To take any of the aforementioned routes for bias mitigation, it is critical to audit the unfairness of the outcome of the predictive algorithms and identify the most severe unfairness issues to address.

Following these concerns, fairness audits of algorithmic decision systems have been pioneered in a variety of fields [13,14]. The auditing process of unfairness detection of a model provides a comprehensive guideline for education researchers and officials to evaluate the inequalities of predictive modeling algorithms from different perspectives before deploying them in practice.

In this paper, we first study if predictive modeling techniques for student success show inequalities for or against a sample of marginalized communities. We use a real national-level education dataset to analyze cases of discrimination. We consider a wide range of ML models for student success prediction. Then, we audit if the prediction outcomes are discriminating against certain subgroups considering different notions of fairness to identify a potential bias in the predictions. Furthermore, we investigate the impact of imputing the missing values using various techniques on model performance and fairness to provide key insights for educational practitioners for responsible ML pipelines. This study has the potential to significantly impact the practice of data-driven decision making in higher education by investigating the impact of a critical pre-processing step on predictive inequalities. In particular, we examine how imputation impacts the performance and fairness of a student success prediction outcome.

Furthermore, we present a prospective evaluation of predictive modeling on operational data as an important step in assessing the real-world performance of ML models. We conduct a prospective evaluation for out-of-sample performance evaluation, whereby a model learned from historical data is evaluated by observing its performance on new data not identically nor independently distributed (iid) as the historical data. Prospective evaluation is likely to provide a less biased estimation of future performance than the evaluation of historical data. Ensuring that a model continues to perform well when integrated into decision-making workflows and provides trustworthy operational impact, we simulate testing observations by adding subtle noise to a different group of predictors and generate non-iid testing sets. To this end, we perturb the sensitive attribute, *race*; non-sensitive attributes; and all predictors in three different scenarios to demonstrate the external generalization capacity to the community.

We predict the most common proxy attribute, *bachelor's degree (B.Sc.) completion*, concerning equal treatment to different demographic groups through different notions of fairness. The comprehensive study of the real large-scale dataset of ESL allows us to validate the performance of different ML techniques for predictive analytics in higher education in a real situation. To the best of our knowledge, none of the existing fair ML models have studied existing large-scale datasets for student success modeling. Most of the extant applications of fair ML demonstrate results using small datasets considering a limited number of attributes (e.g., [12,15]) or in a specific context, such as law school admission [16,17].

2. Bias in Education

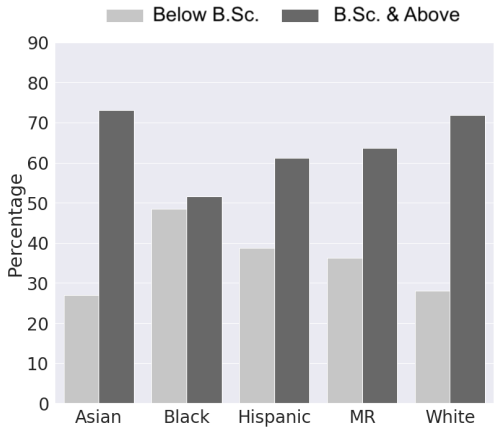
"Bias in, bias out". The first step towards auditing and addressing disparity in student success prediction is to understand and identify different sources of bias in the dataset. Most of the social data, including education data, are almost always biased since they inherently reflect historical biases and stereotypes [18]. Data collection and representation methods often introduce additional bias. Disregarding the societal impact of modeling, the biased data worsens the discrimination in the outcomes of predictive modeling.

The term bias refers to demographic disparities in the sampled data that compromise its representativeness [18,19]. *Population bias* in the data prevents a model from being accurate for minorities [20]. Table 1 presents the racial population bias in the ELS dataset (the Education Longitudinal Study (ELS:2002) is a nationally representative study of 10th graders in 2002 and 12th graders in 2004 https://nces.ed.gov/surveys/els2002/ accessed on 5 January 2024). Note that the dataset used for this analysis is limited to students at 4-year institutions. Therefore, students who identify as "White" are considered as the majority accounting for 66% of the observations, while "Multiracial" (we refer to students with two or more races as multiracial and denote it as MR), "Hispanic" and "Black" groups are underrepresented, making them minorities. On the other hand, bias also exists in the distribution of attribute values across different demographic groups, which is referred to as *behavioral bias*. Such bias in the data has a direct impact on the algorithmic outcomes. One explanation is the strong correlation between sensitive attributes with other attributes.

Table 1. The ELS population bias.

Race	Percent of the Population
Asian	11.13%
Black	10.21%
Hispanic	8.13%
Multiracial	4.27%
White	65.83%

Figure 1a highlights the frequent occurrence of degree attainment below the bachelor's level among Black and Hispanic communities, indicating behavioral bias. Similarly, Figure 1b reveals a lower degree attainment among students from middle- and low-income families, indicating another instance of behavioral bias. These findings emphasize the need for careful consideration and action to mitigate the impact of population and behavioral biases when using degree attainment as a student success indicator.



(**a**)

Family Income	Attainment	% of the Pop.
Lower Class	Below B.Sc.	8.3%
	B.Sc. and Above	10.7%
Middle Class	Below B.Sc.	18.0%
	B.Sc. and Above	37.7%
Upper Class	Below B.Sc.	5.2%
	B.Sc. and Above	20.1%

(**b**)

Figure 1. ELS behavioral bias: (**a**) Degree attainment; (**b**) Attainment vs. family income.

To further explore potential sources of behavioral bias, Figure 2a–c present racial disparities in total credits earned, math/reading test scores, and GPA. Figure 2a demonstrates that the median earned credits for Black and Hispanic groups are lower, as indicated by the lower values of their first and second quartiles. Similarly, Figure 2b shows that the median standardized combined math/reading test score is lower for Black and Hispanic groups. Moreover, Figure 2c reveals a lower median GPA for vulnerable students (Black and Hispanic). The varying size of the boxplots provides insights into the distribution of outcomes within each group. Notably, the boxplots for Black and Hispanic subgroups are larger, indicating greater diversity in total credits earned compared to the more concentrated distribution around the median observed for the White subgroup in Figure 2a. These findings shed light on disparities in educational outcomes and highlight the need for targeted interventions to address behavioral biases and promote equitable opportunities for all students.

By uncovering bias within the available data and recognizing the characteristics of vulnerable populations, this study emphasizes the potential for discrimination to be amplified through predictive modeling. It highlights the critical role of addressing and acknowledging bias during the initial data pre-processing phase to mitigate discriminatory outcomes. In light of this, our investigation aims to examine the influence of data pre-processing, specifically the imputation of missing values, on disparities in prediction outcomes. The detection of bias guides the selection of appropriate data pre-processing approaches to mitigate the negative impact of imputation on model performance and fairness. We conducted an analysis of unfairness in predictive outcomes before and after imputation, considering various scenarios that can arise in real-world applications. Through this analysis, we demonstrate the variations in unfairness results and provide an in-depth discussion of the rationale and implications behind each strategy.

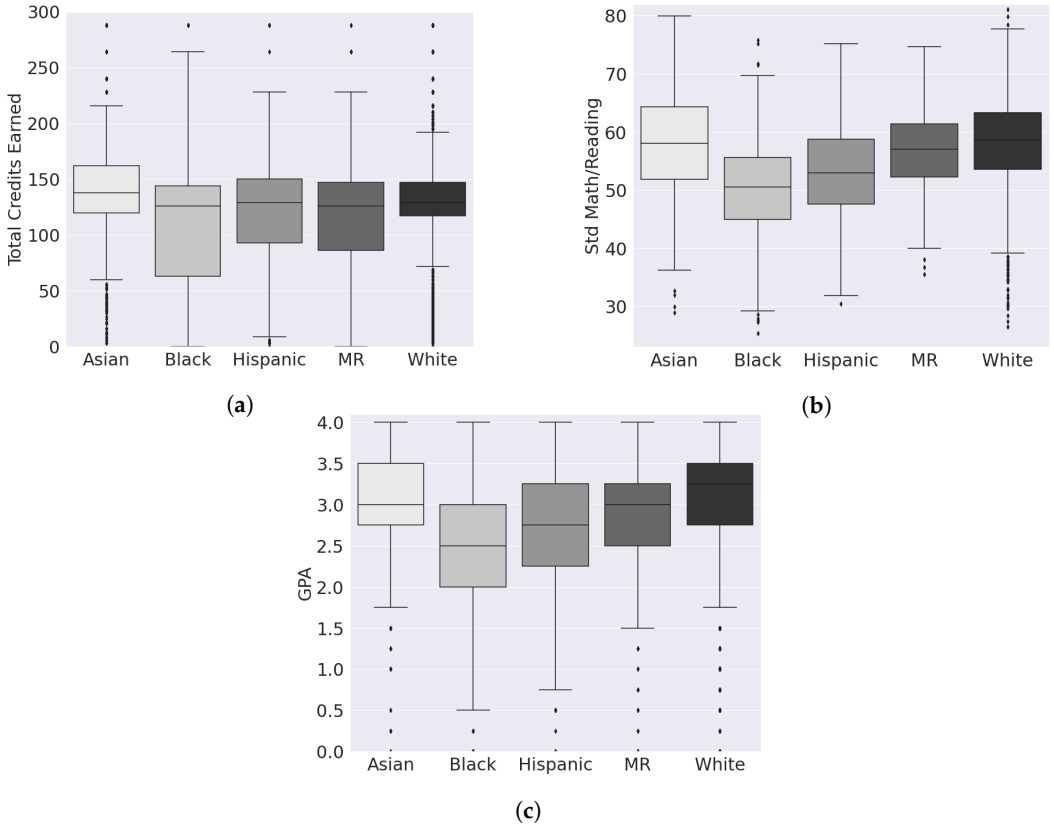


Figure 2. Boxplots of the unprotected attributes associated with $\not\perp$ race for: (**a**) Total credits earned; (**b**) Std Math/Reading score; (**c**) GPA.

3. Fairness in Predictive Modeling

Fairness-aware learning has received considerable attention in the ML literature (fairness in ML) [21,22]. More specifically, fairness in ML seeks to develop methodologies such that the predicted outcome becomes fair or non-discriminatory for individuals based on their protected attributes such as race and sex. The goal of improving fairness in learning problems can be achieved by intervention in pre-processing, in-processing (algorithms), or post-processing strategies. Pre-processing strategies involve the fairness measure in the data preparation step to mitigate the potential bias in the input data and produce fair outcomes [23–25]. In-process approaches [26–28] incorporate fairness into the design of the algorithm to generate a fair outcome. Post-process methods [23,29,30] manipulate the outcome of the algorithm to mitigate the unfairness of the outcome of the decision-making process.

Evaluating the fairness of algorithmic predictions requires a notion of fairness, which can be difficult to choose in practice. There are various definitions for fairness in the literature [2,19,21,31–33] that fall into different categories including Statistical Parity [34], Equalized Odds [34], Predictive Equality [35] and Equal Opportunity [36]. Table 2 demonstrates the mathematical definitions of each of these common metrics [37].

Table 2. Common fairness definitions.

Fairness Notion	Formulation
Statistical Parity (**SP**)	$P(\hat{Y}=1\|S=1)-P(\hat{Y}=1\|S=0)$
Equalized Odds (**EO**)	$P(\hat{Y}=1\|Y=y,S=1)-P(\hat{Y}=1\|Y=y,S=0),\forall y\in\{0,1\}$
Equal Opportunity (**EOP**)	$P(\hat{Y}=1\|Y=1,S=1)-P(\hat{Y}=1\|Y=1,S=0)$
Predictive Equality (**PE**)	$P(\hat{Y}=1\|Y=0,S=1)-P(\hat{Y}=1\|Y=0,S=0)$

Let $S = \{0,1\}$ be a binary sensitive attribute; in a binary classification setting, let $Y = \{0,1\}$ be the true label, and let $\hat{Y} = \{0,1\}$ be the predicted class label. Most of the fairness notions are derived based on conditional probabilities using these variables to reveal the inequalities of the predictive model. Different metrics have been leveraged regarding different contexts, business necessities, and regulations. A predictive modeling outcome might have inequalities under one notion of fairness and might not have any under the others.

The choice of notions and the evaluation of model fairness hold significant significance within educational settings [33,38]. One approach is outlined in the study by Gardner et al. [33], wherein the authors propose a method to evaluate unfairness in predictive models using a metric called ABROCA (Absolute Between-ROC Area). ABROCA enables the comparison of model performance across various subgroups by measuring the differential accuracy between them based on the ROC curve. This framework offers a quantitative means to assess how predictive models may inadvertently favor or disproportionately impact different student subgroups. However, our analysis encompasses a broader range of unfairness quantification, incorporating commonly recognized fairness notions. Details regarding these notions can be found in Table 2. It is worth noting that the fairness evaluation and guidelines derived from our selected metrics can also be extended to ABROCA, a combination of the false positive rate and true positive rate, making it compatible with the same evaluation framework used for the other metrics. To give some examples in the education context, (a) demographic (statistical) parity is referred to as the *discrepancy of the positive prediction of bachelor's degree attainment ($\hat{Y} = 1$) across different demographic groups of students*, and (b) Equal Opportunity indicates the *discrepancy of the positive prediction of bachelor's degree attainment ($\hat{Y} = 1$) across different demographic groups of students, given their success at attaining a bachelor's degree ($Y = 1$)*. In this paper, we use a binary classification setting but across multilevel racial population subgroups (S is not necessarily binary). We extend the fairness metrics, as described in Table 2, for non-binary sensitive attributes by considering a *one-versus-rest* approach for unfairness calculation. More specifically, to calculate the unfairness gaps, we consider each subgroup as $S = 1$ and compare it against the rest $S = 0$ (i.e., all other subgroups), one at a time. We mainly focus on racial disparities; however, our proposed approach for auditing fairness and investigating the imputation impact can be extended to use other sensitive attributes. For example, the decision-maker can use "gender" as a sensitive attribute.

Despite the growing awareness of biases and unfairness in ML, practitioners face numerous challenges, as highlighted in previous studies [39,40], which often focus on specific contexts such as predictive policing [9] and child mistreatment detection [41]. These challenges revolve around the difficulties ML practitioners encounter when applying existing auditing and de-biasing methods to their particular contexts [40]. Consequently, there has been a recent surge in the concept of auditing algorithms and ethics-based auditing in various contexts [13,42–44]. The ultimate goal of the fairness auditing process is to determine the fairness of the results produced by ML models. This process helps identify appropriate actions to address biases, select suitable bias mitigation methods, and determine the most appropriate techniques to employ throughout the ML development pipeline [43].

When it comes to predictive modeling in higher education, it is crucial to support designated users in auditing the performance of ML models and assessing any resulting inequalities before adopting and deploying them in practice. Therefore, in this paper, we aim to assist education practitioners and policymakers in evaluating the inequalities present in predictive outcomes. We conduct fairness audits of ML models used for student success prediction, employing key notions of fairness to identify algorithmic biases using different metrics (Table 2). Additionally, we perform fairness audits to ensure an ethical data pre-processing approach. Utilizing the ELS dataset, our analysis encompasses a wide range of fairness metrics and provides a comprehensive examination of the performance

of different ML models and their associated inequalities across various racial and gender subgroups throughout the data preparation (imputation) and model training steps.

Furthermore, we acknowledge the significance of conducting model assessments under both non-iid and iid settings for prospective evaluations. Non-iid evaluations take into account scenarios where future observations may deviate from the current distribution, capturing the dynamic nature of real-world educational environments. By incorporating non-iid evaluations, we gain insights into the robustness of the ML models and their fairness under different distributional shifts. This broader perspective enhances our understanding of the models' performance and potential biases, enabling more informed decision making.

4. Handling Missing Values in Education

Missing values are common causes of data analysis challenges, affecting modeling, prediction, accuracy, and fairness, especially for protected (sensitive) subgroups. *Handling missing values* is a complex problem that requires careful consideration in education research [3,4]. In the literature, various imputation techniques have been proposed, and the effectiveness of each methodology on different applications has been studied. Here, we briefly describe well-known imputation strategies, such as *Mean Imputation*, *Multiple Imputation*, and clustering-based imputation methods, like *KNN-imputation*.

Many large-scale and nationally representative education datasets, including ELS, suffer from a significant number of incomplete responses from research participants. While features with more than 75% missing or unknown values are typically uninformative, most features suffer from less than 25% missing values based on Table 3 and are worth keeping. Removing all observations with missing values induces significant information loss in success prediction.

Table 3. List of variables.

Variables	% Missing	Variables	% Missing
Student-Teacher relationship	33.32	Std Math/Reading	0.02
F3-loan-owed	25.33	F3_Employment	0
%white teacher	23.69	F3_Highest level of education	0
%Black teacher	19.85	High school attendance	0
%Hispanic teacher	17.72	Family Composition	0
TV/video (h/day)	14.87	Race_Hispanic, race specified	0
Work (h/week)	12.06	F3_Separated no partner	0
F2_College entrance	9.75	F3_Never Married w partner	0
Generation	7.06	F3_Never Married no partner	0
F3_GPA (first attended)	6.79	F3_Married	0
F3_GPA (first year)	6.78	F3_Divorced/Widowed w partner	0
F1_TV/video (h/day)	6.76	F3_Divorced/Widowed no partner	0
F1_units in math	6.13	Race_White	0
Athletic level	5.39	Race_More than one race	0
F1_frequency of computer use	4.27	Race_Hispanic, no race specified	0
Total credits earned	4.04	School Urbanicity	0
F1_Std Math	3.64	Race_Black or African Amer	0
First-year credits earned	3.48	Race_Asian, Hawaii/Pac. Isl	0
F3_GPA (all)	3.33	Race_Amer. Indian/Alaska	0
F3_ total credits earned in Math	3.01	Gender_Male	0
F3_total credits earned in Science	2.90	Gender_Female	0
F1_Work (h/week)	2.77	Parents education	0
Homework (h/week)	1.85	Family income level	0
Number of school activities	0.84	F1_Drop out status	0
Native English speaker	0.02	F3_Separated w partner	0

Simple Imputation is a basic imputation strategy that replaces missing observations with the mean (or median) of the available values for the same variable. This method is known to decrease the standard error of the mean but can increase the risk of statistical tests failing to capture the true reality [3,45].

Multiple Imputation (MI) [46] is an advanced imputation strategy that aims to estimate the natural variation in the data by performing multiple missing data

imputations. MI generates a set of estimates through multiple imputed datasets and combines them into a single set of estimates by averaging across different values. The standard errors of parameter estimates produced using this method have been shown to be unbiased [46].

KNN Imputation is a non-parametric imputation strategy that has shown success in various contexts [47]. KNN imputation replaces missing values in each sample with the mean value from the K nearest neighbors found in the dataset. Two samples are considered close neighbors if the features that they do have are similar. KNN imputation can capture the underlying structure in the dataset even when the data distribution is unknown [48]. To the best of our knowledge, KNN imputation has not been extensively explored in the education context.

Ignoring missing data is not an effective approach to handling missing values, and more importantly, it can result in predictive disparity for minority groups. While many education-related studies have addressed the challenges of missing data, as discussed, little is known about the impact of different imputation techniques on fairness outcomes in the final model. This project aims to address this gap by considering the three aforementioned imputation strategies.

5. Experiments

5.1. The Student Success Prediction Case Study

Before delving into the ML pipeline, we first address the prediction problem at hand. The focus of this paper is specifically on predicting the academic success of students in higher education. Student success prediction is critical for institutional performance evaluation, college admissions, intervention policy design, and various other use cases within higher education [12,49].

Quantifying student success is a complex task because the true quality of any candidate is obscured and the available information is limited. Proxy attributes such as first-year GPA or degree completion are often used as measures of success. In this study, our primary interest lies in predicting the highest level of degree (a classification problem) using the ELS dataset. Numerous factors can affect student success [50]. Thus, identifying the most informative and significant subset of potential variables is a critical task in predictive modeling [51]. To select an appropriate subset of attributes, we conducted an extensive literature search and combined it with domain expert knowledge. These factors include, but are not limited to, academic performance indicators (e.g., SAT scores, GPA) [52], student demographic attributes (e.g., race, gender) [50,53], socio-economic status [54,55], environmental factors, and extracurricular activities.

Incorporating protected attributes in the modeling procedure has raised concerns in the fair-ML domain [2]. The predictive outcome depends on the available information and the specific algorithm employed. A model may utilize any feature associated with the outcome, and commonly used measures of model performance and fairness remain largely unaffected. However, in some cases, the inclusion of unprotected attributes, such as residential zip code, may adversely impact both the performance and fairness of a model due to a latent correlation with other protected attributes, such as race. In this paper, we conduct fairness audits of the model and examine the impact of imputation when incorporating the sensitive attribute as a determining factor.

Table 3 provides a list of variables used in this study along with their corresponding percentages of missing values. It is worth noting that the ELS dataset is survey-based, where F1, F2, and F3 refer to the first follow-up (2004), second follow-up (2006), and third follow-up (2012) years, respectively. Other variables are obtained from the base-year (2002) questionnaires.

5.2. Experimental Setup

The ELS dataset contains several categorical variables. Therefore, our initial step involves appropriately labeling and converting categorical attributes to numerical ones using dummy variable encoding, following the documentation provided by the NCES (https://nces.ed.gov/surveys/els2002/avail_data.asp accessed on 5 January 2024).

Subsequently, we transform the target variable, *highest level of degree*, into a binary classification problem. Specifically, we assign a label of 1 to students with a college degree (BS degree and higher), representing the favorable outcome, while we label others as the unfavorable outcome with a label of 0. It is important to note that we filter the dataset to include only students who attended four-year post-secondary institutions based on institution type. For fairness evaluation, we considered "race" as the sensitive attribute, which comprises five groups: *White, Black, Hispanic, Asian*, and *multiracial (MR)* students.

To ensure data quality, we conduct data cleaning procedures. This includes identifying and renaming missing values based on the dataset documentation and removing observations with a high number of missing attributes, defined as those with more than 75% missing attribute values. More importantly, this study makes the assumption of missingness being completely at random (MCAR) based on conducted statistical tests, correlation analysis, and data examination. Consequently, we assume that there are no statistically significant relationships between missingness and observed variables. Nonetheless, it is crucial to acknowledge that the assumption of MCAR is made to facilitate analysis and cannot be definitively proven [56].

We consider three imputation techniques as discussed in Section 4: Simple Imputation **SI**, Multiple Imputation **MI**, and KNN Imputation **KNN-I**. In addition, we include a baseline approach where we remove observations with missing attributes, which we refer to as **Remove-NA**. By comparing the performance and fairness outcomes of these different imputation methods, we aim to assess their impact on the predictive models and fairness measures.

Following the data preparation step and obtaining clean-format datasets, we proceed with the model training procedure.

Our objective is to analyze the performance of various ML models under each imputation technique and testing scenario, with a specific focus on auditing the inequalities in the prediction outcomes. In our analysis, we incorporate the following prominent machine learning (ML) models within the realm of higher education: Decision Tree (**DT**) [57], Random Forest (**RF**) [58], Support Vector Classifier **SVC** [59], and Logistic Regression (**Log**) [60].

5.3. Prospective Evaluation

Prospective evaluation plays a vital role in the context of ML in education, especially when predicting student success. It allows us to assess the performance and fairness of predictive models on new and unseen data that simulate real-time deployment scenarios. By conducting prospective evaluations, we gain insights into how well the models can generalize to future data and how they may perform in real-world educational settings to address important research questions regarding the effectiveness and fairness of ML models in educational contexts. The findings from this evaluation will provide valuable insights for researchers, educators, and policymakers interested in leveraging ML for student success prediction.

In our study, we aim to investigate the impact of imputation and data representation on the accuracy and fairness of the predictive modeling outcomes, conducting a comprehensive and trustworthy evaluation using various scenarios to assess the generalization capacity of the models. These scenarios enable us to assess the models' generalization capacity and their ability to handle different data situations. Table 4 summarizes the scenarios considered in our evaluation.

Table 4. Evaluation scenarios.

Scenario	Missing Values	Train Test Data
RNA.rnd	Removed	80:20 split, iid
RNA.str	Removed	80:20 split, stratified on race and response variable, iid
Imp.rnd	Imputed	80:20 split, iid
Imp.str	Imputed	80:20 split, stratified on race and response variable, iid
Imp.prop	Imputed	80:20 split and fixing the fraction of observations within each group to match the entire dataset, iid
RNA.rnd.perturb	Imputed	RNA.rnd scenario but the testing data is perturbed for non-sensitive attributes, non-iid
Imp.rnd.perturb	Imputed	Imp.rnd scenario but the testing data is perturbed for non-sensitive attributes, non-iid
Imp.prop.perturb	Imputed	Imp.prop scenario but the testing data is perturbed for non-sensitive attributes, non-iid
RNA.rnd.perturb.sensitive	Imputed	RNA.rnd scenario but the testing data is perturbed for the sensitive attributes, non-iid
Imp.rnd.perturb.sensitive	Imputed	Imp.rnd scenario but the testing data is perturbed for the sensitive attributes, non-iid

Before describing the scenarios, it is important to discuss the concept of independent identically distributed (iid) and non-iid data for the prospective evaluation. In iid scenarios, such as **Imp.rnd**, **RNA.rnd**, and **Imp.prop**, the testing data follows the same distribution as the training dataset. This assumption is commonly made in the machine learning literature. On the other hand, in non-iid scenarios, such as **Imp.prop.perturb** and **Imp.rnd.perturb**, the modified testing dataset deviates from the distribution observed in the training data due to the random perturbations applied to the attributes.

Now, let us discuss the evaluation scenarios in detail.

1. **RNA.rnd** scenario: In this scenario, we follow the common practice in the ML literature and remove all rows with missing values before performing a train/test split. This ensures that the testing data follows the same distribution as the training data and serves as a baseline for comparison.
2. **RNA.str** scenario: In this scenario, we remove missing values while ensuring stratification on both the "race" and "response" variables. This guarantees that each racial subgroup and success outcomes are well-represented in both the training and testing datasets, addressing potential biases introduced by missing data.
3. **Imp.rnd** scenario: In this scenario, we split the entire dataset into train/test sets, perform an imputation technique on the training set to replace missing values, and transfer the trained imputer to replace missing values in the testing set. The imputation technique allows us to retain more data and potentially improve predictive performance.
4. **Imp.str** scenario: Similar to the **Imp.rnd** scenario, we perform imputation on the training set and transfer the imputer to replace missing values in the testing set. However, we also consider stratification on the "race" and "response" variables to generate representative training and testing datasets, addressing potential biases related to imputation and ensuring fair evaluation.
5. **Imp.prop** scenario: In this scenario, we aim to maintain the distribution of different racial groups in the train/test splits. We fix the fraction of observations from each racial group in the splits to match the fractions observed in the entire available dataset. This scenario helps assess the impact of imputation on fairness by preserving the representation of different racial subgroups.

Furthermore, to accurately simulate real-time scenarios, we add random perturbations to modify the distribution of attributes within the testing dataset. This approach acknowledges that future observations may not adhere to the same distribution as the training data. We categorize these scenarios as perturbation scenarios. In the perturbation scenarios, including **Imp.prop.perturb**, **Imp.rnd.perturb**, and **RNA.rnd.perturb**, we exclusively apply perturbations to the non-sensitive attributes while leaving the sensitive attributes unaffected. Conversely, in **Imp.rnd.perturb.sensitive** and **RNA.rnd.perturb.sensitive**, we solely perturb the sensitive attributes. In the **Imp.prop.perturb** scenario, it is worth

mentioning that the perturbation of sensitive attributes is not considered. This deliberate omission is to ensure the preservation of the proportion of sensitive sub-populations. By focusing on preserving the original distribution of sensitive attributes, we can specifically examine the impact of perturbations on the non-sensitive attributes while keeping the relative representation of different racial groups intact. This approach allows us to explore the effects of attribute perturbations on model performance and fairness without introducing additional bias or skewing the representation of sensitive sub-populations.

It should be noted that the perturbation scenarios are categorized as non-independent identically distributed (non-iid). In these scenarios, the testing dataset is intentionally modified to deviate from the distribution observed in the training data due to the applied perturbations. By introducing these non-iid scenarios, we aim to assess the models' robustness and adaptability to distribution shifts and varying patterns in student data.

5.4. Discussion and Analysis

In this section, we present our key findings and insights through three main discussions. First, we compare the performance and fairness of different ML models using various imputation strategies. Next, we compare the outcomes of random and stratified imputed data against the **Remove-NA** scenarios. We also explore the impact of imputation on unfairness in different perturbation scenarios, considering unknown distributions for future observations in real-time scenarios.

As shown in Table 2, different fairness metrics provide insights into the unfairness exhibited across racial groups in the student success prediction outcomes. Statistical Parity (**SP**) focuses on the comparison of positive prediction outcomes across racial groups, without considering the true outcome. Predictive Equality (**PE**) examines the unfairness of incorrectly predicting unsuccessful students as successful based on their racial subgroup. Equal Opportunity (**EOP**) focuses on the correct classification of successful students based on their racial subgroup. Equalized Odds (**EO**) measures the true positive and false positive rates across racial groups.

Figure 3 illustrates the effect of employing different imputation techniques on the fairness of ML models, as measured by the Statistical Parity (SP) notion. It is evident that imputation leads to increased unfairness for Black and Hispanic student groups. However, it is important to note that imputation significantly reduces the variance of unfairness across racial groups by incorporating more observations. This observation can be attributed to the significant increase in the size of the testing set after imputation, as shown in Table 5. As we include more observations from the minority (underprivileged) group of students through imputation, the average unfairness gaps widen further. Despite variations in imputation strategies, their adverse impact on unfairness remains consistent across the different ML models. Therefore, the evaluated imputation techniques in this study exhibit similar effects on unfairness. Furthermore, we observe a consistent pattern in the impact of different imputation methods on unfairness across various fairness notions discussed in Appendix A. In addition, upon comparing the accuracy of various machine learning (ML) models in Figure A4 of Appendix A, it is evident that both **MI** and **KNN-I** outperform other imputation techniques, as indicated by their higher average accuracy values. Although imputation improves prediction accuracy by incorporating additional observations for model training, it exacerbates the unfairness for Black and Hispanic groups. Moving forward, **MI** is predominantly employed to impute missing values in the scenarios described in Section 5.5.

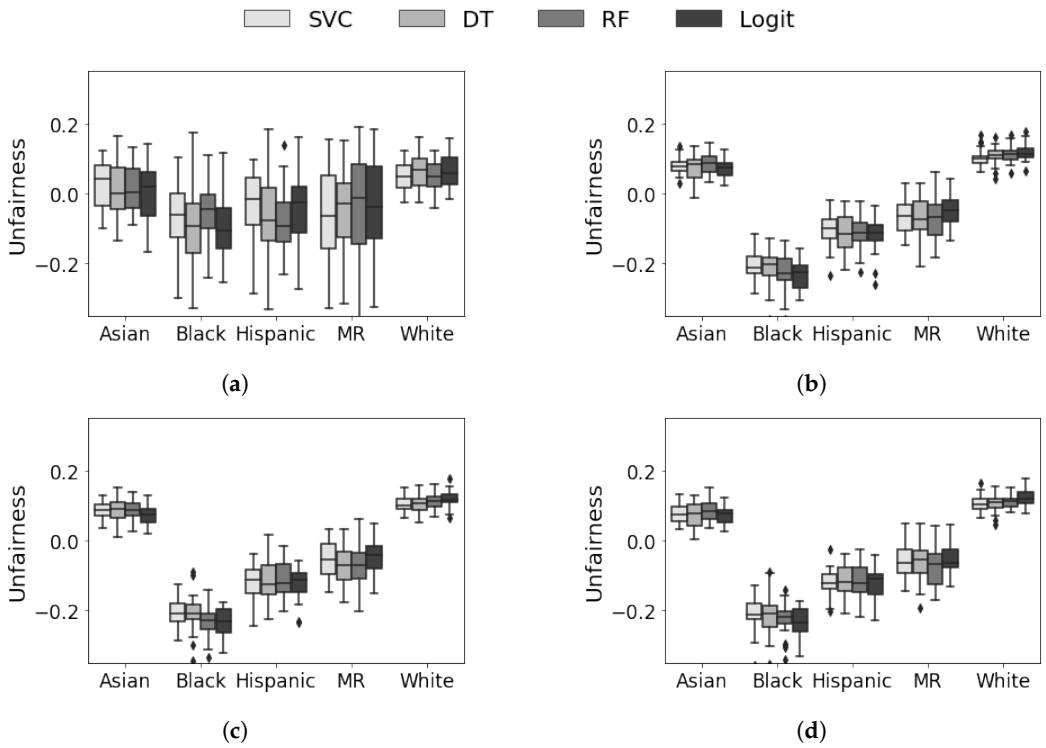
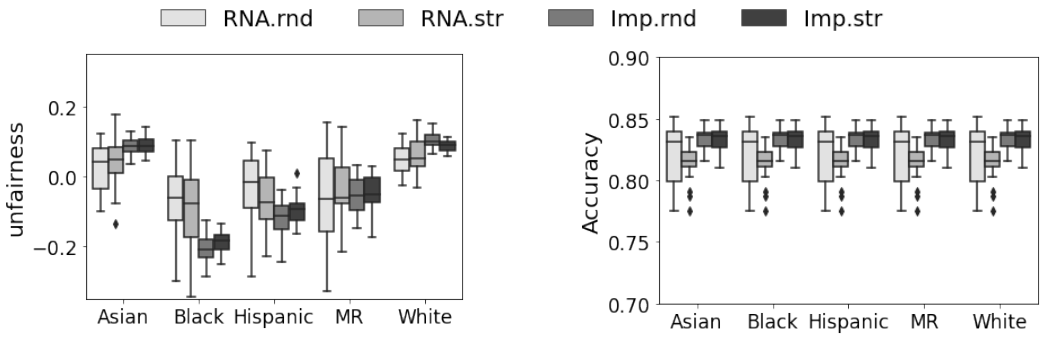


Figure 3. The impact of imputation—Statistical Parity. (**a**) Remove-NA; (**b**) Simple Imputation; (**c**) Multiple Imputation; (**d**) KNN Imputation.

Table 5. Population size of the demographic groups.

Scenario	Race	Pop. X_{train}	Pop. X_{test}	$Y = 1$	$\hat{Y} = 1$
RNA.rnd	Asian	95	26	20	22
	Black	75	19	13	15
	Hispanic	75	17	13	14
	Multiracial	32	9	7	8
	White	712	175	128	145
Imp.rnd	Asian	494	123	90	108
	Black	452	114	62	77
	Hispanic	357	94	58	67
	Multiracial	190	47	30	35
	White	2923	727	526	598
Imp.prop	Asian	494	123	90	100
	Black	452	114	58	66
	Hispanic	361	90	55	65
	Multiracial	190	47	30	36
	White	2920	730	525	614

Figure 4 illustrates the impact of stratification with and without imputation on the performance and fairness based on **SP** for the **SVC** model. The results demonstrate that the combination of stratification and imputation does not have significant effects on the prediction performance or fairness metrics. However, when comparing the scenarios **RNA.str** and **RNA.rnd** (without imputation), stratification leads to worsened unfairness. It is worth noting that similar patterns are observed with other fairness metrics and models, such as **RF**, as demonstrated in Appendix A.



(**a**) (**b**)

Figure 4. The impact of stratification: Remove-NA vs. Imputation for Support Vector Classifier (SVC). (**a**) Statistical Parity; (**b**) Testing accuracy.

Figure 5 illustrates the performance and fairness of **SVC** for scenarios involving the perturbation of the sensitive attribute "race". Specifically, it compares the outcomes between the non-iid and iid with imputation, **Imp.rnd.perturb.sensitive**, and **Imp.rnd**, as well as the non-iid and iid without imputation, **RNA.rnd.perturb.sensitive** and **RNA.rnd**. Our empirical analysis uncovered that when significant changes occur in the distribution of future observations based solely on their sensitive attribute (i.e., changes in racial group representation), the removal of missing values (**RNA.rnd.perturb.sensitive**) can lead to fairer and more accurate predictions for the unprivileged group of students (e.g., Black and Hispanic). However, imputing the missing values results in a significant loss of accuracy and fairness for the newly modified student cohort. Although the average unfairness outcome does not show a significant decline in some cases (e.g., EOP) with imputation, the considerable variation observed in the final unfairness values raises concerns.

One possible explanation is that the perturbation of the sensitive attribute "race" in the non-iid scenarios introduces variations in the distribution of the data. When missing values are removed, the predictive models are not influenced by imputed values that may introduce biases or distortions. As a result, the models can rely solely on the available data, including other relevant features, to make predictions. This approach can lead to fairer outcomes by avoiding potential biases associated with imputation. On the other hand, imputing missing values may introduce additional uncertainty and potential biases in the data. Imputation methods attempt to fill in missing values based on patterns observed in the available data. However, in scenarios where the distribution of future observations changes significantly based on sensitive attributes, imputing missing values may introduce inaccuracies and distortions in the imputed data. This can result in a loss of accuracy and fairness, particularly for the newly modified student cohort. The observed variations in unfairness values with imputation could be attributed to the challenges of accurately imputing missing values in scenarios with significant distribution shifts. Imputation methods may struggle to capture the complex patterns and changes introduced by the perturbations, leading to inconsistent and potentially biased imputations. These inconsistencies can manifest as variations in the unfairness of the outcome, indicating the sensitivity of imputation to changes in the underlying data distribution. The findings indicate that the impact of imputation and perturbation on fairness and accuracy is consistent across different machine learning models, as is shown in Appendix A.

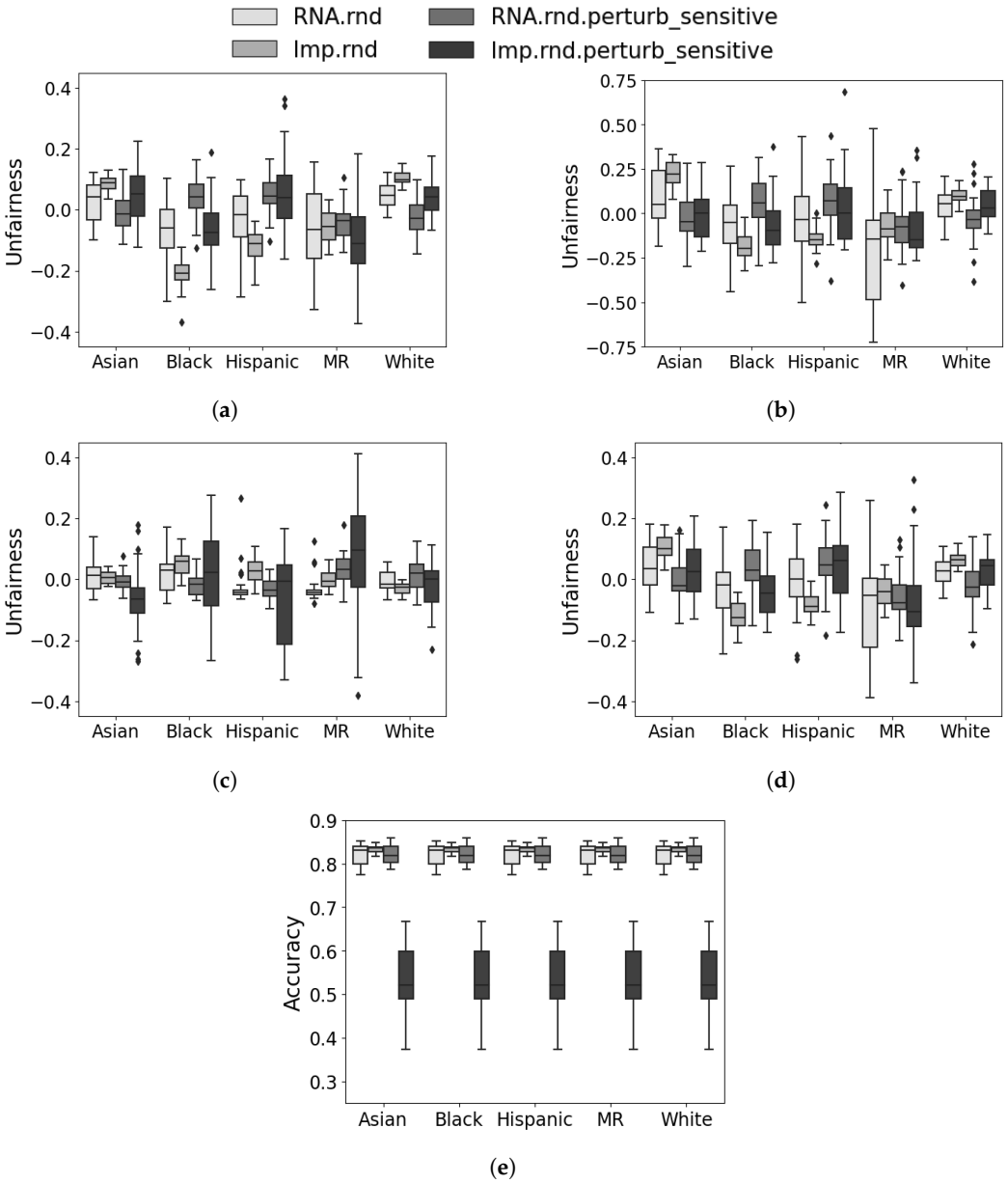


Figure 5. Impact of perturbing sensitive attribute: Remove-NA vs. Imputation for Support Vector Classifier (SVC). (**a**) Statistical Parity; (**b**) Predictive Equality; (**c**) Equal Opportunity; (**d**) Equalized Odds; (**e**) Testing accuracy.

Figure 6 presents the results of various fairness metrics across different evaluation scenarios, specifically focusing on perturbations (non-iid) applied to "non-sensitive" attributes. The results indicate that the **RNA.rnd** scenario exhibits relatively higher fairness compared to the **Imp.rnd** and **Imp.prop** scenarios, particularly concerning minority groups such as Black and Hispanic students. This suggests that removing missing values could potentially result in less biased predictions compared to imputation approaches, assuming that the future batch of students follows a similar iid distribution. However, in perturbation scenarios, both **RNA.rnd.perturb** and **Imp.rnd.perturb** demonstrate comparable average fairness, but with a larger variance than the iid scenarios. Specifically, imputation in the **Imp.rnd.perturb** scenario shows lower unfairness according to several metrics, including the **SP** metric. On the other hand, the removal of missing values in the **RNA.rnd.perturb** scenario emerges as a more favorable strategy for disadvantaged groups based on the **EOP** metric. Additionally, non-iid perturbation scenarios lead to decreased testing accuracy

compared to random iid scenarios due to significant differences in the distribution of non-sensitive predictive attributes.

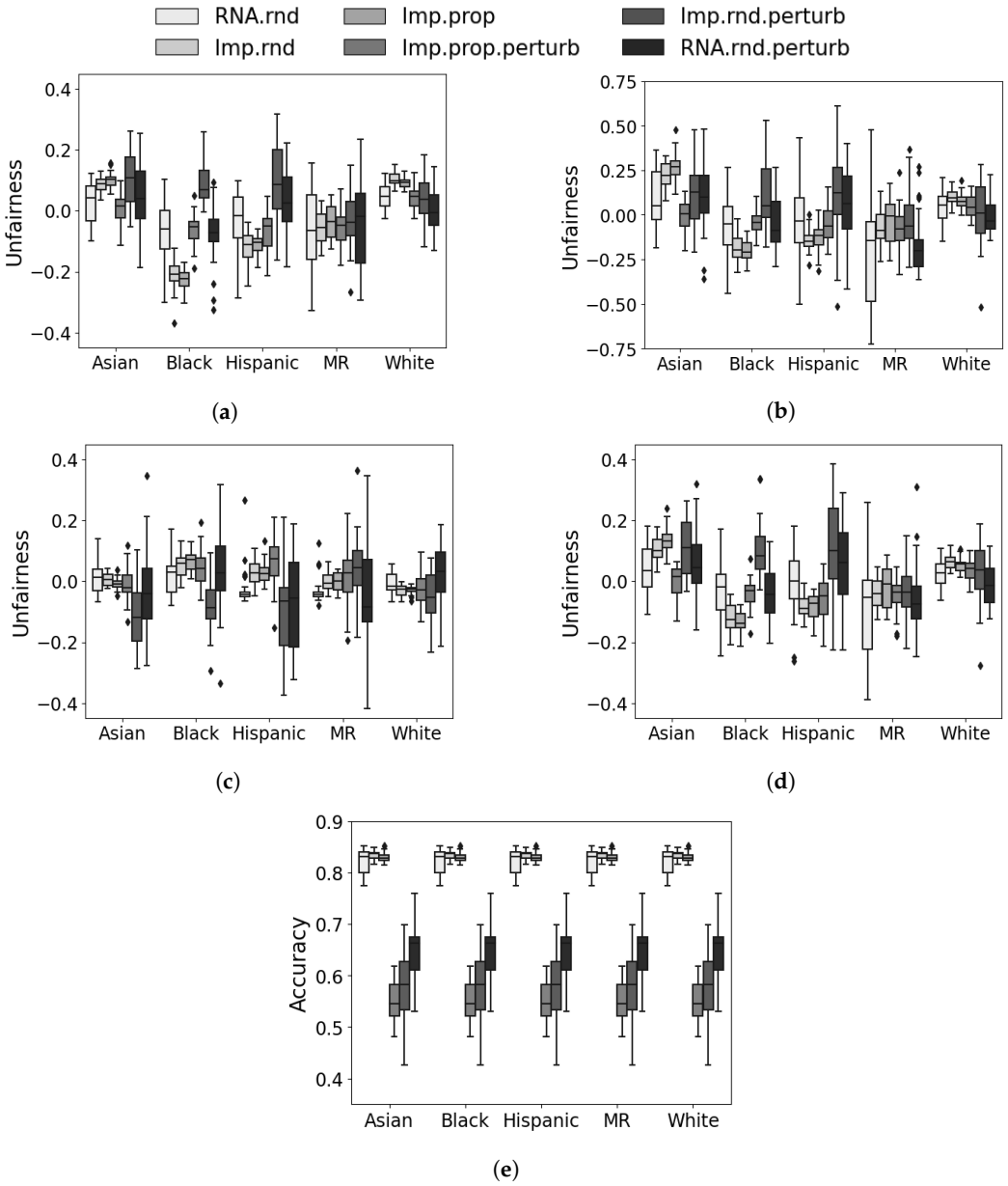


Figure 6. The impact of perturbing non-sensitive attributes: Remove-NA vs. Imputation for Support Vector Classifier (SVC). (**a**) Statistical Parity; (**b**) Predictive Equality; (**c**) Equal Opportunity; (**d**) Equalized Odds; (**e**) Testing accuracy.

The results in Figures 5a and 6a highlight that for **SP**, perturbing either the sensitive attribute (race) or other non-sensitive attributes in the testing data leads to a decrease in the unfairness for Black and Hispanic students. For **PE**, Figures 5b and 6b show that imputation solely based on race and imputation, with consideration of the intervention (perturbing non-sensitive attributes), leads to an almost similar level of unfairness for Black and Hispanic students. Figures 5c and 6c illustrate that perturbing the imputed data leads to a reduction in fairness based on **EOP** for Black and Hispanic students (unprivileged group). The underlying rationale stems from the effect of perturbation on the model's accuracy. Altering the distribution diminishes the likelihood of accurately classifying successful students (true positives), resulting in fewer correct identifications of success. For **EO**, Figures 5d and 6d reveal that when perturbing either the sensitive attribute race or

other non-sensitive attributes in the testing data (in a real-time scenario), there is a tendency for the Black and Hispanic groups to receive fewer positive predictions (regarding success). This reduction in fairness serves as an indication to decision-makers, highlighting the need for policy changes to be taken into account in the future.

These findings collectively imply that choosing between imputation and missing value removal methods should be guided by the desired fairness outcomes and the trade-offs in predictive accuracy given a context. In an ideal situation, where social injustices have been mitigated through effective interventions, the perturbation results showcased in Figure 6 exemplify the potential outcomes of these interventions that result in a different distribution of predictors for different subgroups, emphasizing the importance of addressing societal disparities and implementing interventions for disadvantaged populations. By understanding and addressing these disparities, we can work towards fostering fairness and equity in predictive modeling and decision-making processes.

5.5. Summary of Results

Based on the results discussed here, several key interpretations can be made regarding the impact of missing data imputation and perturbation on fairness and predictive performance in educational predictive modeling scenarios.

Missing data imputation: Imputing missing values can improve prediction accuracy by incorporating more observations into the modeling process. However, it can also exacerbate unfairness, particularly for minority groups. The choice of the imputation method does not significantly impact fairness outcomes, indicating that the evaluated imputation techniques are similar in terms of their impact on fairness. Multiple Imputation (MI) is a recommended approach due to its performance and fairness characteristics.

Stratification: Stratifying the data based on sensitive attributes, such as race, can worsen unfairness, even in scenarios without imputation. This suggests that stratification alone may not be sufficient to mitigate fairness issues and may require additional considerations or modifications.

Perturbation: (a) When perturbing the sensitive attribute "race" in non-iid scenarios, removing missing values (RNA) can lead to fairer and more accurate predictions for underprivileged groups (e.g., Black and Hispanic students). However, imputing missing values results in a significant loss of accuracy and fairness for the modified student cohort. The impact of perturbation on fairness and accuracy varies depending on the specific fairness metrics considered. (b) When perturbing non-sensitive attributes, the impact on fairness and predictive performance becomes more nuanced. Both the removal of missing values (RNA) and imputation (Imp) approaches show comparable average fairness levels. Imputation in the Imp.perturb scenario demonstrates lower unfairness according to various fairness metrics, including the Statistical Parity (SP) metric. On the other hand, the removal of missing values in the RNA.perturb scenario emerges as a more favorable strategy for disadvantaged groups based on the Equal Opportunity (EOP) metric.

These results suggest that in non-iid perturbation scenarios, imputation may provide some advantages in reducing unfairness across the entire dataset. However, the removal of missing values still offers benefits in terms of fairness, specifically for underprivileged groups. It is important to note that non-iid perturbation scenarios lead to decreased testing accuracy compared to random iid scenarios due to significant differences in the distribution of non-sensitive predictive attributes.

Trade-offs: There is a trade-off between fairness and accuracy in educational predictive modeling. While removing missing values can lead to higher fairness levels, it may sacrifice prediction accuracy, especially in perturbation scenarios. Imputing missing values can improve accuracy but may introduce unfairness, particularly for minority groups.

These findings emphasize the complexity of addressing missing data and perturbation in educational predictive modeling. Careful consideration should be given to the trade-offs between fairness and accuracy, and the specific goals and requirements of the educational context should guide the selection of appropriate strategies.

6. Conclusions

Education researchers and decision-makers encounter numerous challenges, including data preparation, when utilizing predictive modeling to inform decision-making processes. Additionally, deploying machine learning on historical data can introduce bias into future predictions. One significant source of bias stems from the representation of vulnerable population subgroups in the historical data. In this study, we examined the impact of handling missing values through imputation on prediction unfairness using the large-scale national ELS dataset.

First, we compared the standard practice of removing all rows with missing values to the process of imputing missing values during the regular train/test split. In comparison to the scenario where missing values are removed, imputation increases the average unfairness gap for vulnerable student groups while reducing its variance. However, imputation also improves the prediction accuracy of various machine learning models, indicating a trade-off with fairness. Furthermore, we found that the choice of imputation strategy has a negligible impact on the unfairness gaps among different population subgroups.

To perform a prospective evaluation for the unfairness of the ML outcomes, we designed various train/test split scenarios and manipulated the distribution of future observations by perturbing different attributes in the testing data. Our analysis demonstrated that while imputation tends to increase the average unfairness in the standard train/test split approach, it aids in achieving fairer models when the distribution of future observations changes (due to effective interventions addressing disparities in society). In other words, if access to educational resources, parental education, community training, and interventions supporting vulnerable students improve, the upcoming cohort of students would exhibit a different distribution. Consequently, a model trained on imputed training data would be fairer. The rationale behind this is that imputation enables the retention of hundreds of additional observations, which can be valuable in scenarios where the future distribution deviates from the patterns observed in historical data.

Future research can focus on conducting a detailed analysis exploring different unfairness mitigation techniques and evaluating their implications on the outcomes. This could involve developing algorithms that explicitly address bias and its impact on predictions, especially concerning vulnerable population subgroups. Additionally, a promising research direction involves a detailed investigation into quantifying the delicate balance between fairness and accuracy within predictive modeling. Through comprehensive analysis employing various fairness metrics across diverse datasets, studies can offer invaluable insights and guidelines. Such investigations aim to shed light on how enhancements in fairness might impact model accuracy, and conversely, how prioritizing accuracy might affect fairness, thereby providing essential guidelines for model development and deployment in the education sector.

Author Contributions: Conceptualization, H.A.; Methodology, H.A. and N.N.; Software, N.N. and P.H.; Validation, H.A., D.G., N.N. and P.H.; Formal analysis, H.A., N.N. and P.H.; Investigation, H.A., N.N. and P.H.; Data curation, N.N. and P.H.; Writing—original draft, H.A.; Writing—review & editing, H.A.; Visualization, N.N. and P.H.; Supervision, H.A. and D.G.; Funding acquisition, H.A. and D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by the Institute of Education Sciences (R305D220055).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is a publicly available dataset that has been used in the literature. This dataset is deidentified and available on the IES website as cited in the main text of the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

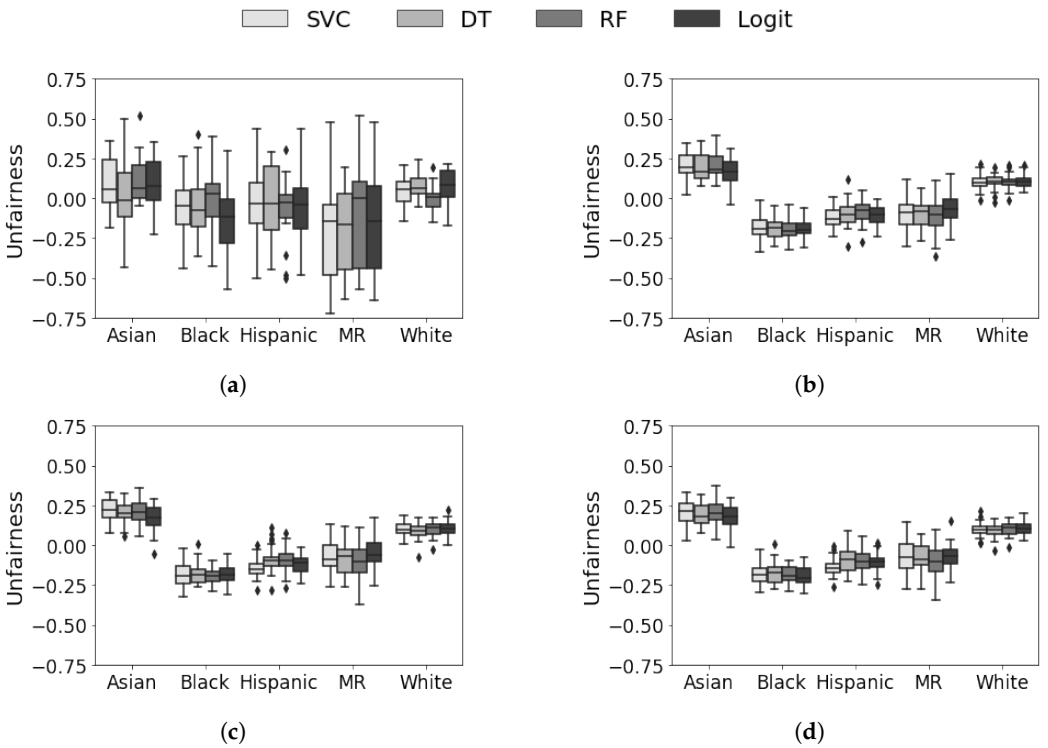


Figure A1. The impact of imputation—Predictive Equality. (**a**) Remove-NA ; (**b**) Simple Imputation; (**c**) Multiple Imputation; (**d**) KNN Imputation.

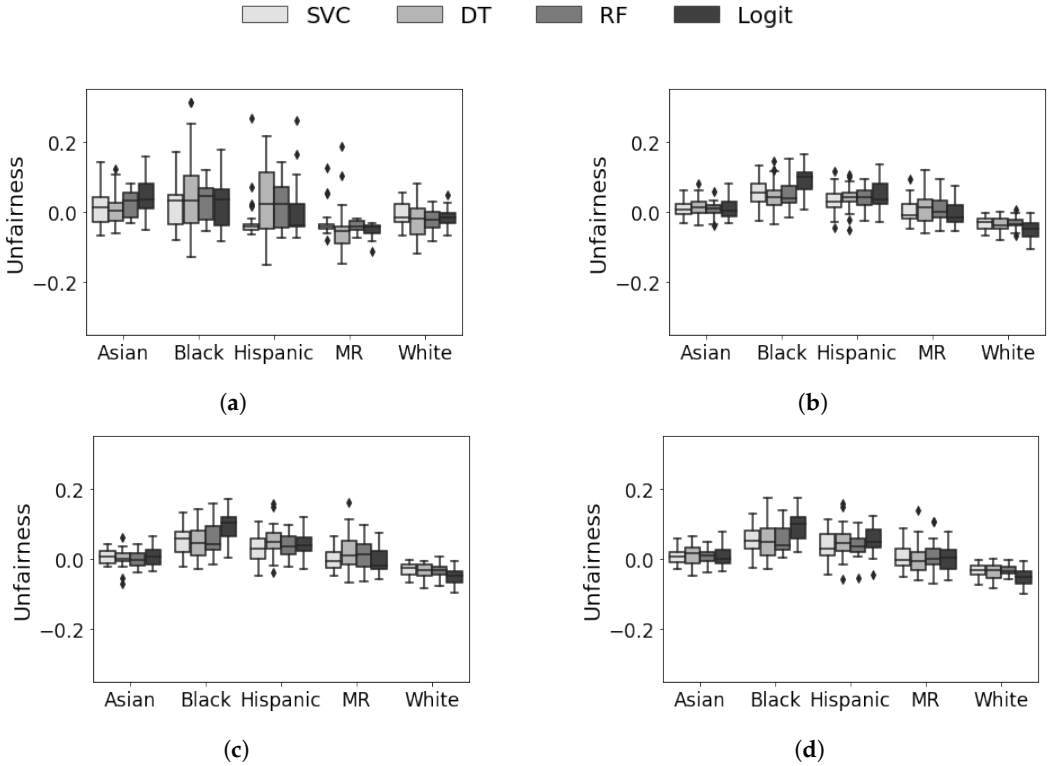


Figure A2. The impact of imputation—Equal Opportunity. (**a**) Remove-NA; (**b**) Simple Imputation; (**c**) Multiple Imputation; (**d**) KNN Imputation.

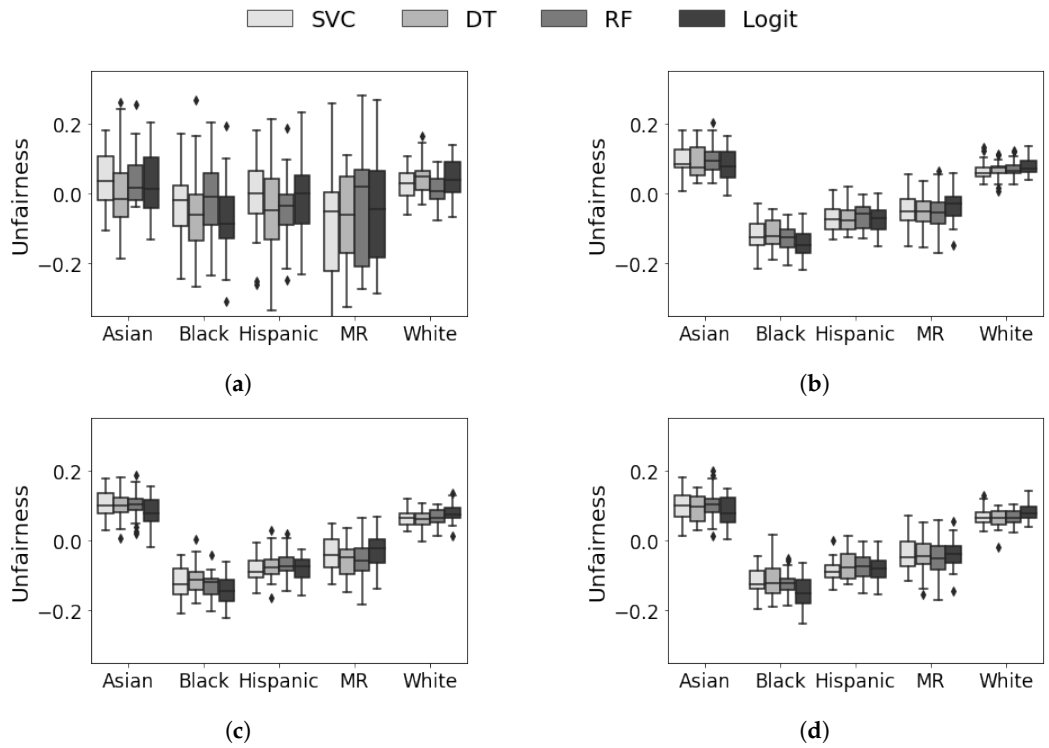


Figure A3. The impact of imputation—Equalized Odds. (**a**) Remove-NA; (**b**) Simple Imputation; (**c**) Multiple Imputation; (**d**) KNN Imputation.

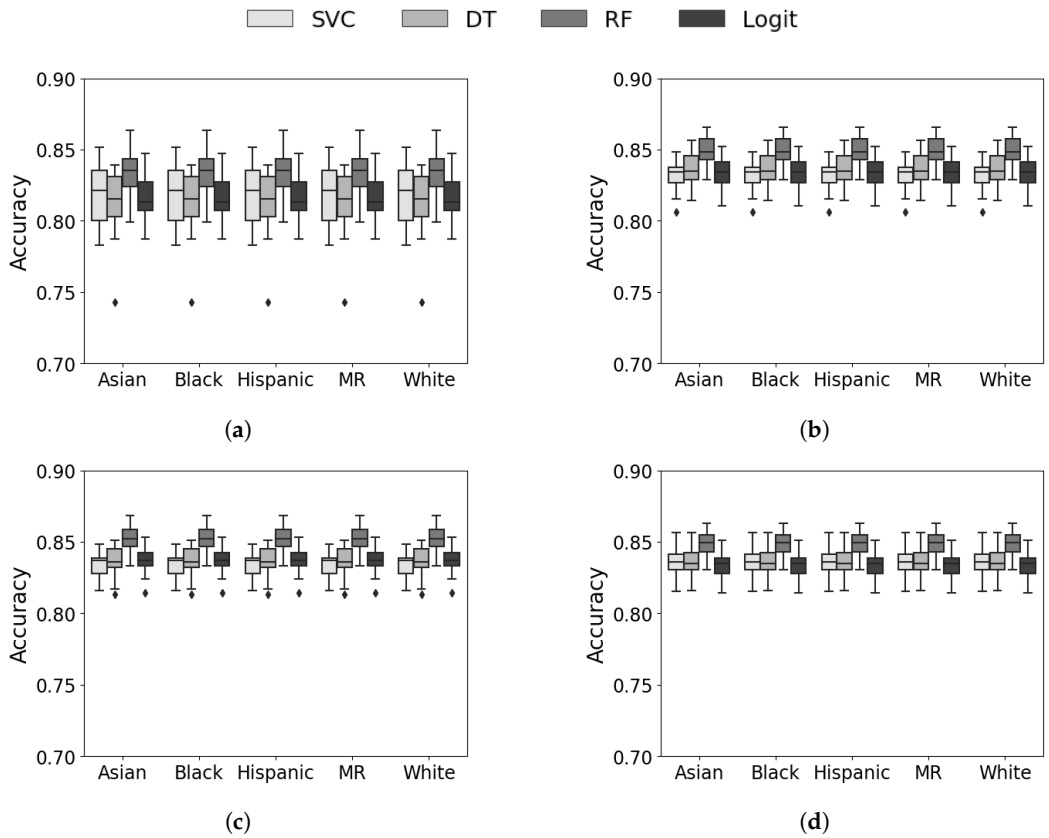


Figure A4. The impact of imputation—Testing accuracy. (**a**) Remove-NA; (**b**) Simple Imputation; (**c**) Multiple Imputation; (**d**) KNN Imputation.

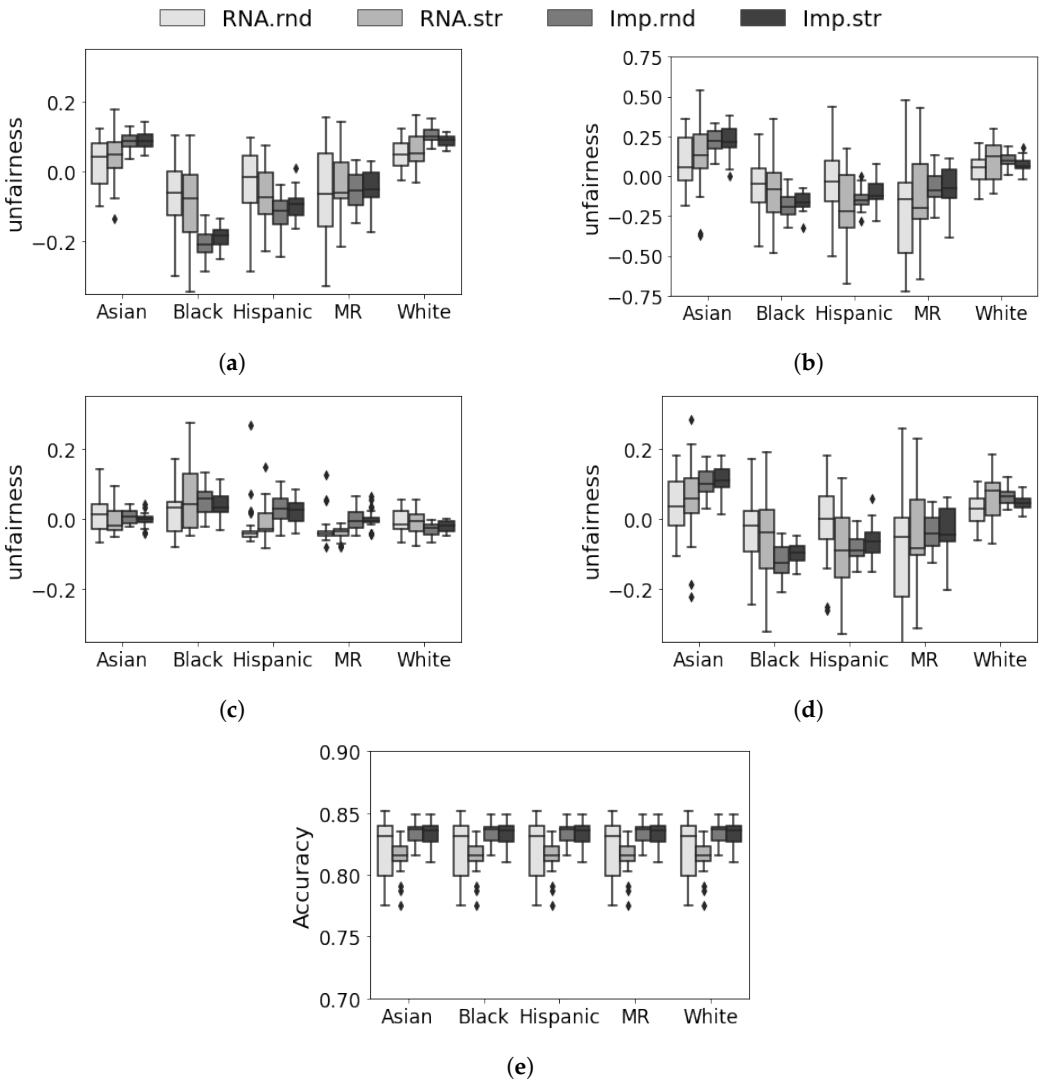


Figure A5. The impact of stratification: Remove-NA vs. Imputation for Support Vector Classifier (SVC). (**a**) Statistical Parity; (**b**) Predictive Equality; (**c**) Equal Opportunity; (**d**) Equalized Odds; (**e**) Testing accuracy.

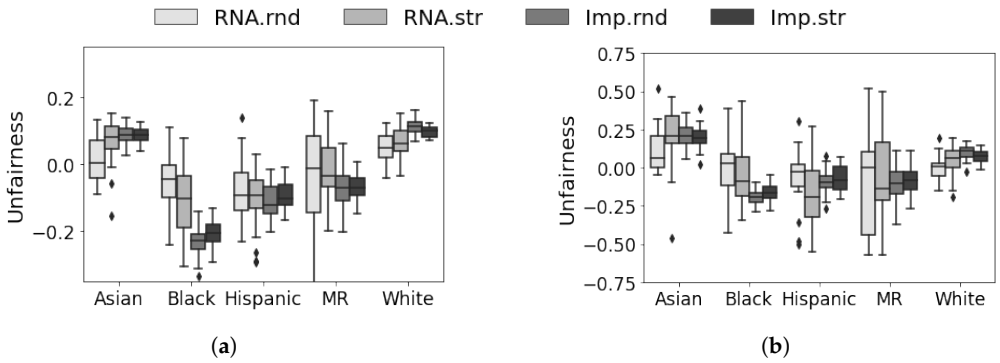


Figure A6. *Cont*.

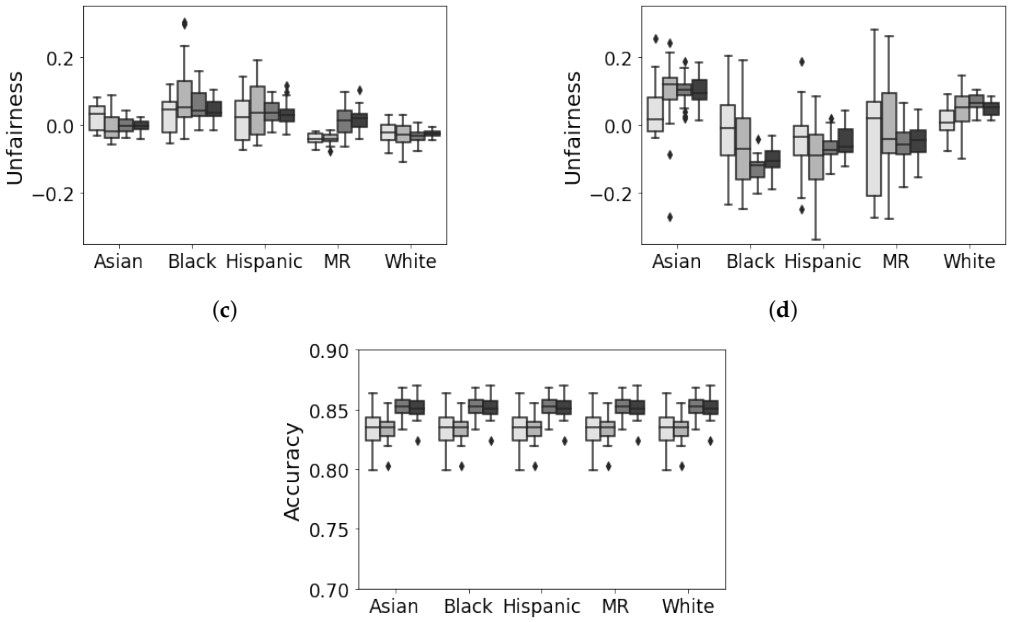


Figure A6. The impact of stratification: Remove-NA vs. Imputation for Random Forest. (**a**) Statistical Parity; (**b**) Predictive Equality; (**c**) Equal Opportunity; (**d**) Equalized Odds; (**e**) testing accuracy.

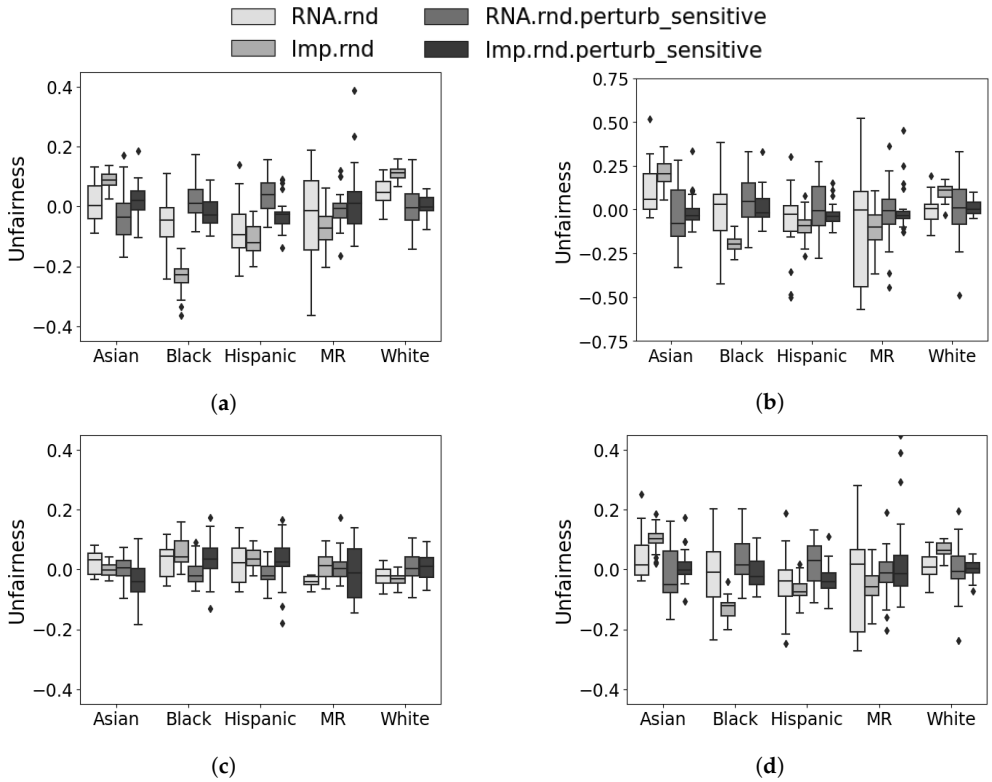
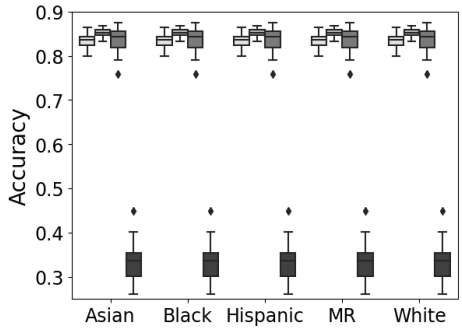
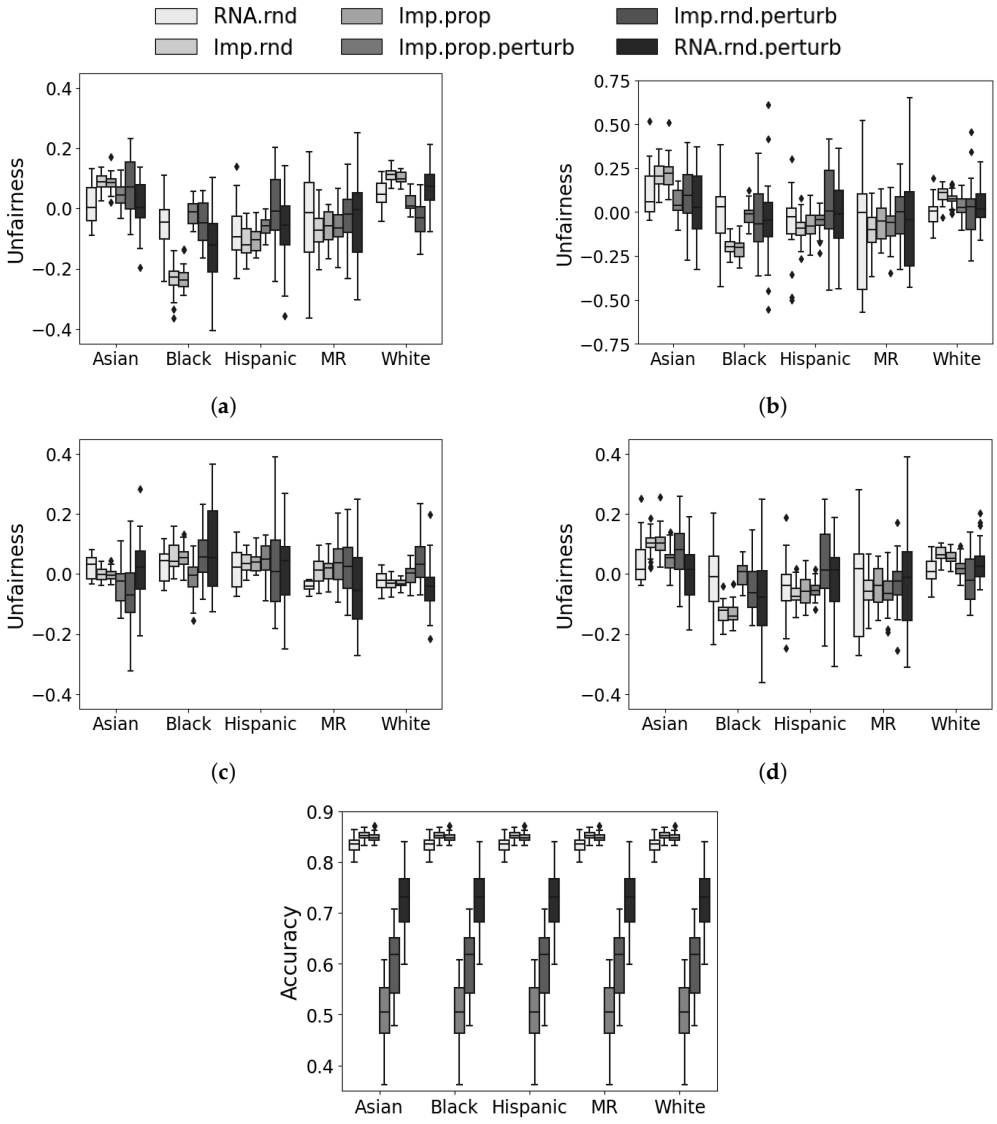


Figure A7. *Cont*.



(**e**)

Figure A7. The impact of perturbing sensitive attributes: Remove-NA vs. Imputation for Random Forest (RF). (**a**) Statistical Parity; (**b**) Predictive Equality; (**c**) Equal Opportunity; (**d**) Equalized Odds; (**e**) testing accuracy.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure A8. The impact of perturbing non-sensitive attributes: Remove-NA vs. Imputation for Random Forest (RF). (**a**) Statistical Parity; (**b**) Predictive Equality; (**c**) Equal Opportunity; (**d**) Equalized Odds; (**e**) testing accuracy.

References

1. Ekowo, M.; Palmer, I. The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis. *New America*, 24 October 2016.
2. Barocas, S.; Selbst, A.D. Big data's disparate impact. *Calif. Law Rev.* **2016**, *104*, 671. [CrossRef]
3. Cheema, J.R. A review of missing data handling methods in education research. *Rev. Educ. Res.* **2014**, *84*, 487–508. [CrossRef]
4. Manly, C.A.; Wells, R.S. Reporting the use of multiple imputation for missing data in higher education research. *Res. High. Educ.* **2015**, *56*, 397–409. [CrossRef]
5. Kwak, S.K.; Kim, J.H. Statistical data preparation: Management of missing values and outliers. *Korean J. Anesthesiol.* **2017**, *70*, 407. [CrossRef] [PubMed]
6. Valentim, I.; Lourenço, N.; Antunes, N. The Impact of Data Preparation on the Fairness of Software Systems. In Proceedings of the 2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE), Berlin, Germany, 28–31 October 2019; pp. 391–401.
7. Fernando, M.P.; Cèsar, F.; David, N.; José, H.O. Missing the missing values: The ugly duckling of fairness in machine learning. *J. Intell. Syst.* **2021**, *36*, 3217–3258. [CrossRef]
8. Kizilcec, R.F.; Lee, H. Algorithmic Fairness in Education. *arXiv* **2020**, arXiv:2007.05443.
9. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias: Risk Assessments in Criminal Sentencing. *ProPublica* **2016**. Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 5 January 2024).
10. Feathers, T. Major Universities Are Using Race as a "High Impact Predictor" of Student Success. 2021. Available online: https://themarkup.org/news/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success (accessed on 5 January 2024).
11. Marcinkowski, F.; Kieslich, K.; Starke, C.; Lünich, M. Implications of AI (un-) fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice and organizational reputation. In Proceedings of the ACM FAccT, Barcelona, Spain, 27–30 January 2020.
12. Yu, R.; Li, Q.; Fischer, C.; Doroudi, S.; Xu, D. Towards accurate and fair prediction of college success: Evaluating different sources of student data. In Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020), Virtual, 10–13 July 2020; pp. 292–301.
13. Kondmann, L.; Zhu, X.X. Under the Radar—Auditing Fairness in ML for Humanitarian Mapping. *arXiv* **2021**, arXiv:2108.02137.
14. Kearns, M.; Neel, S.; Roth, A.; Wu, Z.S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2564–2572.
15. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Rambachan, A. Algorithmic fairness. In *AEA Papers and Proceedings*; American Economic Association: Nashville, TN, USA, 2018; Volume 108, pp. 22–27.
16. Kusner, M.J.; Loftus, J.R.; Russell, C.; Silva, R. Counterfactual fairness. *arXiv* **2017**, arXiv:1703.06856.
17. Cole, G.W.; Williamson, S.A. Avoiding resentment via monotonic fairness. *arXiv* **2019**, arXiv:1909.01251.
18. Olteanu, A.; Castillo, C.; Diaz, F.; Kiciman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* **2019**, *2*, 13. [CrossRef]
19. Barocas, S.; Hardt, M.; Narayanan, A. Fairness and Machine Learning: Limitations and Opportunities. 2019. Available online: http://fairmlbook.org (accessed on 5 January 2024).
20. Asudeh, A.; Jin, Z.; Jagadish, H. Assessing and remedying coverage for a given dataset. In Proceedings of the ICDE, Macao, China, 8–11 April 2019; pp. 554–565.
21. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the ITSC, Anchorage, AK, USA, 16–19 September 2012; pp. 214–226.
22. Zafar, M.B.; Valera, I.; Rogriguez, M.G.; Gummadi, K.P. Fairness constraints: Mechanisms for fair classification. In Proceedings of the Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 962–970.
23. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In Proceedings of the SIGKDD, Sydney, Australia, 10–13 August 2015; ACM: New York, NY, USA, 2015; pp. 259–268.
24. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33. [CrossRef]
25. Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.N.; Varshney, K.R. Optimized pre-processing for discrimination prevention. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3992–4001.
26. Zafar, M.B.; Valera, I.; Rodriguez, M.G.; Gummadi, K.P. Fairness constraints: Mechanisms for fair classification. *arXiv* **2015**, arXiv:1507.05259.
27. Zhang, H.; Chu, X.; Asudeh, A.; Navathe, S.B. OmniFair: A Declarative System for Model-Agnostic Group Fairness in Machine Learning. In Proceedings of the SIGMOD, Xi'an, China, 20–25 June 2021; pp. 2076–2088.
28. Anahideh, H.; Asudeh, A.; Thirumuruganathan, S. Fair active learning. *arXiv* **2020**, arXiv:2001.01796.
29. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K.Q. On fairness and calibration. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5680–5689.
30. Zehlike, M.; Bonchi, F.; Castillo, C.; Hajian, S.; Megahed, M.; Baeza-Yates, R. Fa* ir: A fair top-k ranking algorithm. In Proceedings of the CIKM, Singapore, 6–10 November 2017; pp. 1569–1578.

31. Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.* **2017**, *31*, 1060–1089. [CrossRef]
32. Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In Proceedings of the ACM FAT*, New York, NY, USA, 23 August 2018.
33. Gardner, J.; Brooks, C.; Baker, R. Evaluating the fairness of predictive student models through slicing analysis. In Proceedings of the 9th International Conference on Learning Analytics & Knowledge, Tempe, AZ, USA, 4–8 March 2019; pp. 225–234.
34. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
35. Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq, A. Algorithmic decision making and the cost of fairness. In Proceedings of the SIGKDD, Halifax, NS, Canada, 13–17 August 2017; ACM: New York, NY, USA, 2017; pp. 797–806.
36. Madras, D.; Creager, E.; Pitassi, T.; Zemel, R. Fairness through causal awareness: Learning causal latent-variable models for biased data. In Proceedings of the ACM FAT*, New York, NY, USA, 17 January 2019; pp. 349–358.
37. Makhlouf, K.; Zhioua, S.; Palamidessi, C. On the applicability of machine learning fairness notions. *ACM SIGKDD Explor. Newsl.* **2021**, *23*, 14–23. [CrossRef]
38. Anahideh, H.; Nezami, N.; Asudeh, A. On the choice of fairness: Finding representative fairness metrics for a given context. *arXiv* **2021**, arXiv:2109.05697.
39. Veale, M.; Van Kleek, M.; Binns, R. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the ACM CHI, Montreal, QC, Canada, 21–26 April 2018; pp. 1–14.
40. Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the ACM CHI, Glasgow, UK, 4–9 May 2019; pp. 1–16.
41. Chouldechova, A.; Benavides-Prado, D.; Fialko, O.; Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In Proceedings of the ACM FAT*, New York, NY, USA, 23–24 February 2018; pp. 134–148.
42. Mokander, J.; Floridi, L. Ethics-based auditing to develop trustworthy AI. *arXiv* **2021**, arXiv:2105.00002.
43. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the ACM FAccT, Barcelona, Spain, 27–30 January 2020; pp. 33–44.
44. Wilson, C.; Ghosh, A.; Jiang, S.; Mislove, A.; Baker, L.; Szary, J.; Trindel, K.; Polli, F. Building and auditing fair algorithms: A case study in candidate screening. In Proceedings of the ACM FAccT, Online, 3–10 March 2021; pp. 666–677.
45. Allison, P.D. *Missing Data*; Sage Publications: New York, NY, USA, 2001.
46. Rubin, D.B. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* **1996**, *91*, 473–489. [CrossRef]
47. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525. [CrossRef]
48. Somasundaram, R.; Nedunchezhian, R. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. *Int. J. Comput. Appl.* **2011**, *21*, 14–19. [CrossRef]
49. Stephan, J.L.; Davis, E.; Lindsay, J.; Miller, S. *Who Will Succeed and Who Will Struggle? Predicting Early College Success with Indiana's Student Information System*; REL 2015-078; Regional Educational Laboratory Midwest: Washington, DC, USA, 2015.
50. Voyer, D.; Voyer, S.D. Gender differences in scholastic achievement: A meta-analysis. *Psychol. Bull.* **2014**, *140*, 1174. [CrossRef] [PubMed]
51. Ramaswami, M.; Bhaskaran, R. A study on feature selection techniques in educational data mining. *arXiv* **2009**, arXiv:0912.3924.
52. Chamorro-Premuzic, T.; Furnham, A. Personality, intelligence and approaches to learning as predictors of academic performance. *Personal. Individ. Differ.* **2008**, *44*, 1596–1603. [CrossRef]
53. Zlatkin-Troitschanskaia, O.; Schlax, J.; Jitomirski, J.; Happ, R.; Kühling-Thees, C.; Brückner, S.; Pant, H. Ethics and fairness in assessing learning outcomes in higher education. *High. Educ. Policy* **2019**, *32*, 537–556. [CrossRef]
54. Filmer, D. *The Structure of Social Disparities in Education: Gender and Wealth*; The World Bank: Washington, DC, USA, 2000.
55. Avdic, D.; Gartell, M. Working while studying? Student aid design and socioeconomic achievement disparities in higher education. *Labour Econ.* **2015**, *33*, 26–40. [CrossRef]
56. Heitjan, D.F.; Basu, S. Distinguishing "missing at random" and "missing completely at random". *Am. Stat.* **1996**, *50*, 207–213.
57. Hamoud, A.; Hashim, A.S.; Awadh, W.A. Predicting student performance in higher education institutions using decision tree analysis. *Int. J. Interact. Multimed. Artif. Intell.* **2018**, *5*, 26–31. [CrossRef]
58. Pelaez, K. Latent Class Analysis and Random Forest Ensemble to Identify At-Risk Students in Higher Education. Ph.D. Thesis, San Diego State University, San Diego, CA, USA, 2018.
59. Agaoglu, M. Predicting instructor performance using data mining techniques in higher education. *IEEE Access* **2016**, *4*, 2379–2387. [CrossRef]
60. Thompson, E.D.; Bowling, B.V.; Markle, R.E. Predicting student success in a major's introductory biology course via logistic regression analysis of scientific reasoning ability and mathematics scores. *Res. Sci. Educ.* **2018**, *48*, 151–163. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.