Rubric-based holistic review represents a change from traditional graduate admissions approaches in physics

Nicholas T. Young,[1,2,*] N. Verboncoeur,[1] Dao Chi Lam,[3] and Marcos D. Caballero[1,2,4,5,†]

[1]*Department of Physics and Astronomy, Michigan State University, East Lansing, Michigan 48824, USA*
[2]*Department of Computational Mathematics, Science, and Engineering, Michigan State University, East Lansing, Michigan 48824, USA*
[3]*Department of Statistics, Michigan State University, East Lansing, Michigan 48824, USA*
[4]*Center for Computing in Science Education and Department of Physics, University of Oslo, N-0316 Oslo, Norway*
[5]*CREATE for STEM Institute, Michigan State University, East Lansing, Michigan 48824, USA*

(Received 13 December 2021; revised 15 November 2022; accepted 13 April 2023; published 15 May 2023)

Rubric-based admissions are claimed to help make the graduate admissions process more equitable, possibly helping to address the historical and ongoing inequities in the U.S. physics graduate school admissions process that have often excluded applicants from minoritized races, ethnicities, genders, and backgrounds. Yet, no studies have examined whether rubric-based admissions methods represent a fundamental change in the admissions process or simply represent a new tool that achieves the same outcome. To address that, we developed supervised machine learning models of graduate admissions data collected from our department over a seven-year period. During the first four years, our department used a traditional admission process and switched to a rubric-based process for the following three years, allowing us to compare which parts of the applications were driving admissions decisions. We find that faculty focused on applicants' physics GRE scores and grade-point averages when making admissions decisions before the implementation of the rubric. While we were able to develop a sufficiently good model whose results we could trust for the data before the implementation of the rubric, we were unable to do so for the data collected after the implementation of the rubric, despite multiple modifications to the algorithms and data such as implementing Tomek Links. Our inability to model the second dataset despite being able to model the first combined with model comparison analyses suggests that rubric-based admission does change the underlying process. These results suggest that rubric-based holistic review is a method that could make the graduate admission process in physics more equitable.

DOI: 10.1103/PhysRevPhysEducRes.19.010134

I. INTRODUCTION

While graduate school has historically been seen as a route for students to begin careers in academia, graduates are increasingly pursuing careers across industry, government, and academia. The National Science Foundation's Survey of Doctorate Recipients finds that less than half of all Ph.D.s work at an educational institution while only 2 out of 5 physics Ph.D.s do [1]. As such, universities have a duty to ensure that their students are able to achieve their chosen career trajectories.

Yet, the data suggest that is not always the case. Only three out of five physics students who enroll in a Ph.D. program will successfully complete their program [2,3]. As undertaking graduate study involves a significant time and financial investment from both the student and institution, failing to ensure students graduate is a waste of resources. Solutions must consider both the admission and retention sides of this problem. In this paper, we will focus on the former.

As the Council of Graduate Schools notes in one of its reports, "Better selection [of graduate students] can result in higher completion rates" [4]. Historically and continuing today, graduate school admissions in the United States have tended to be an exclusionary process that favors certain groups over others. Previous research into the graduate admissions process in physics has found that the process relies heavily on the quantitative metrics such as grade point average (GPA) and general and physics GRE scores [5–10]. These metrics have been found to benefit groups already overrepresented in higher education. For example,

*ntyoung@umich.edu
†Corresponding author.
caball14@msu.edu

prior work has shown students from groups underrepresented in higher education (e.g., first generation, low income, Black, Latinx, Native) suffered a grade penalty relative to their more privileged peers with students from minoritized racial groups suffering the largest penalties [11]. Other work has shown that the general and physics GREs are biased against women and students from minoritized racial and ethnic groups [2,12] as well as against students from smaller or less prestigious universities [13]. Furthermore, the high costs associated with these often-required tests, despite limited evidence that these tests serve a predictive purpose [2,14,15], prevent some students from applying [16,17].

The inequities in the admissions process and the fact that traditional admissions methods "miss many talented students" [18] have led various programs and organizations to consider alternative admission approaches such as holistic admissions, which considers a "broad range of candidate qualities including 'noncognitive' or personal attributes" [19]. These efforts are often supported by rubrics to ensure that all applicants are assessed on the same explicit criteria, provide a structure to do so, and reduce implicit bias present in the admissions process [8,20].

The physics department at Michigan State University was one such program that has taken this approach. Beginning with applicants applying for entry to the program in Fall 2018, all applicants were evaluated on a 18-construct rubric covering areas such as academic performance, research experience, noncognitive skills, community contributions, and standardized test scores. Our initial analysis of the changes to our graduate admissions process suggests that the rubric appears to reduce implicit bias and that switching to this rubric-based holistic review process increased the percent of women enrolling in our program [21]. These results were consistent with analyses of rubric-based admissions in related disciplines like engineering [22,23].

However, it is difficult to know whether these rubrics are changing the underlying inequitable admissions process currently in use or if they are merely new tools that continue to center inequitable components of the application in admissions decisions. For example, even in departments actively working to increase their diversity, prior work has found that GPA and GRE scores still had an undue influence on who was admitted [9].

Therefore, thinking about how to address inequities in graduate admissions in physics, we ask how does the introduction of a rubric change a program's admissions process? We then operationalize this question into two research questions:

1. What parts of a graduate application determine whether an applicant will be admitted to a physics program?
2. How does the introduction of a rubric with defined constructs to evaluate applicants affect which parts of the application determine admission?

To answer these questions, we compare admission models of our current graduate admissions process using data from both faculty's ratings of applicants using the rubric and the applications to historical data of our program's initial process. In our initial analysis of the historical data [24], we noticed there are cases where applicants have similar physics GRE scores and GPA, yet one applicant is accepted while the other is not. Given that cases such as these might add challenges to modeling the data, removing such applicants might allow us to better characterize the general trends in the data. We, therefore, consider an alternative approach that detects similar applicants with different admission outcomes and removes them from the dataset: Tomek Links [25]. We then ask a third research question:

3. How does using Tomek Links affect our ability to answer the first two research questions?

Unlike other studies in physics graduate admissions, this work represents a case study of a single institution rather than a broad look at the graduate admissions landscape. However, because physics is regarded as a high consensus discipline, that is, there is large agreement about what counts as legitimate admissions practices [26], we believe our results will be relevant to similar doctoral programs. We do acknowledge that our historical data do not capture all parts of the graduate school application, namely, the written components and that may hurt the generalizability of the study.

II. FRAMING ADMISSIONS AS A COMPUTATIONAL PROBLEM

When evaluating applicants to a graduate program, faculty are presented with information about the applicant and must make a judgment as to whether to admit or reject the applicant. Whether the applicant is admitted or rejected depends on a set of criteria developed by the faculty members reviewing the applicant. As such, we choose to frame the problem of understanding admissions as a classification problem, where a computer must use a set of rules to determine what the qualitative outcome should be or was [27].

We choose to apply a classification machine learning model under this computational framing, specifically random forest [28], instead of a more traditional technique for classifying data such as logistic regression (as used by Attiyeh and Attiyeh [29] and Posselt *et al.* [9] to study graduate admissions and by Barceló *et al.* to study holistic admissions in residency interview screening and selection [30]) due to the lack of assumptions on the data and random forest's feature importances. Because the random forest is not based on an underlying probabilistic model like logistic regression is, it can handle almost any distribution of data and does not require a linear relationship between the outcome and predictor variables. Given that graduate admissions is a complicated process, we should not expect

a linear relationship between the parts of the application and admission status.

Regarding the feature importances, they measure all factors on the same scale, allowing factors of otherwise different scales can be compared. This contrasts with logistic regression where the odds ratio for a continuous variable would measure the change in odds for a unit increase in the variable while the odds ratio for a categorical or binary variable measures the change in odds relative to a reference group. In addition, the feature importances allow for each categorical feature as a whole to be compared to the other features rather than in pairs relative to the reference group. This property can be especially useful for features like an applicant's proposed research area where there is no natural or standard choice of reference group or when we are not interested in category differences. That is, random forest could tell us how much faculty care about what subfield the applicant is interested in while logistic regression would only be able to tell us whether someone in experimental high energy physics is more likely to be admitted than someone in theoretical condensed matter physics.

Here, we will assume that faculty are primarily seeking to admit applicants who are likely to succeed in their graduate program. As Small notes, there are other possibilities such as aligning with research needs or funding, increasing diversity and inclusion, or developing talent from a specific geographic area [31]. We will also assume that these applicants are included in the data but represent only a small fraction of the cases.

When evaluating the potential "success" of an applicant in the program, there will likely be cases where an argument for and against admission can be made. While admissions committees use common criteria for initially judging applicants, deliberations of these borderline applicants under the traditional process might come down to subtle distinctions that were not used for other applicants [8]. Thinking in terms of a modeling perspective, this means that some applicants might be assessed according to additional and potentially implicit criteria and hence, these borderline applicants might not be easily classified by a general model of the admissions process. As a result, including these borderline applicants might cause our model performance to suffer. Alternatively, excluding these applicants and instead focusing on a more typical applicant could improve model performance and provide better insight into whether the underlying process changed.

Unfortunately, whether an applicant is a borderline applicant is not included in faculty ratings of applicants and hence, we do not know who is a borderline applicant. To determine who might be a borderline applicant, let us assume there is a predictive model of a graduate admissions process that perfectly separates those who are admitted and those who are not admitted in some n-dimensional application space. We could then say that those applicants who are near the $n-1$-dimensional boundary that separates the admitted applicants and not admitted applicants are borderline applicants. To differentiate borderline applicants in the admission process from borderline applicants in the modeling process, we will refer to the latter as *boundary applicants*. Such a definition of *boundary applicants* is like Hoens and Chawla's definition of borderline cases in classification, which are cases where a small change in the features would cause the classification boundary to shift [32].

However, such an approach assumes that those who are admitted and not admitted can be cleanly split in some n-dimensional space and are not intermixed. For a variety of reasons (such as those listed in Small [31]), an applicant with a stellar application might be rejected or an applicant with a weaker application might be admitted and hence an admitted applicant might fall on the not-admitted side of the separating boundary or vice versa. While these applicants might not be borderline in the traditional sense, their admission decision likely would have required deliberation and hence, might have gone through a similar process as a borderline applicant. We should therefore also consider these applicants as borderline applicants in the sense of the possibility of hurting our model's performance. Perhaps more accurately, we should refer to these applicants as *noise applicants* following Hoens and Chawla's definition of noise cases, which are cases that result from random variation and are not representative of the underlying pattern [32].

While we have operationalized borderline applicants in terms of a model as *boundary applicants* and *noise applicants*, we still need a method to determine which applicants these are before constructing any models. Tomek Links offers one possible method as it is a method of identifying the boundary or noise cases in the data [25].

To identify the Tomek Links in a dataset, the distances between all cases in the dataset are computed. Using the distances, the nearest neighbor of each case is computed. For two cases, e.g., case 1 and case 2, the cases are Tomek Links if and only if case 1 is the nearest neighbor of case 2, case 2 is the nearest neighbor of case 1, case 1 and case 2 are of different classes. The only way for these conditions to be fulfilled is if case 1 and case 2 are boundary cases or if case 1 or case 2 is a noise case [32]. Therefore, Tomek Links allows us to identify *boundary applicants* and *noise applicants* in our data. An example of this approach in practice is shown in Fig. 1. In theory, removing these *boundary* and *noise* applicants from our datasets should then allow us to create models more representative of the underlying trends. We acknowledge that this process removes cases that might be of potential interest. For example, students with low physics GRE scores who were nevertheless admitted and other cases that go against the norm would likely be removed by Tomek Links.

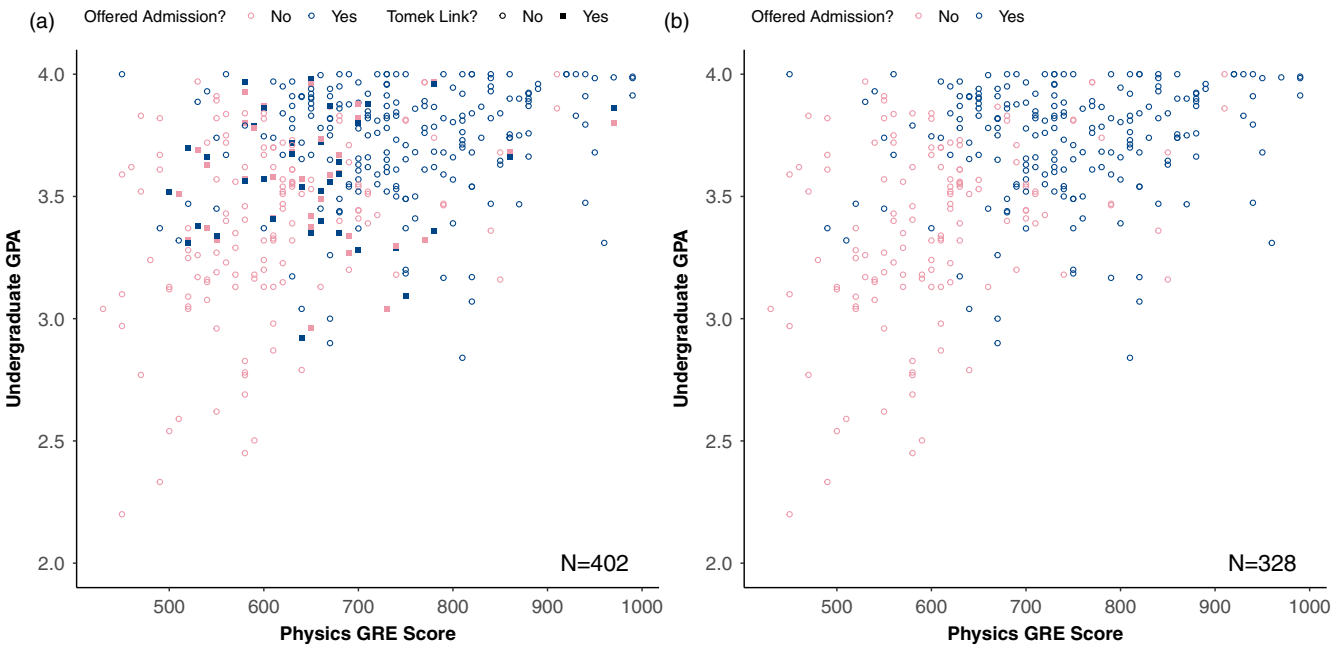

FIG. 1. Plot A shows Fig. 2 of Young and Caballero [24] with the Tomek Links marked. Filled square points represent Tomek Links. Plot B shows the same plot after the Tomek Links have been removed. Data for which either the GPA or physics GRE score is missing are not plotted.

While Tomek Links have been successfully used in other contexts (e.g., see [33–35]), these approaches have tended to use data augmentation in conjunction with Tomek Links. While data augmentation approaches are valid from a modeling perspective, they might be questionable from an ethics and policy perspective. For example, altering the dataset might lead to a model that is highly inaccurate of the underlying process [36]. For our dataset, using data augmentation is analogous to creating applicants and thus our conclusions about how our admissions process might or might not have changed would be based on both real and imaginary applicants. For this reason, we will not use data augmentation.

As we note in our methods, we do impute our data. Readers may view this as a contradiction of the previous paragraph, but we view data imputation and data augmentation as different. Data imputation is using the existing data to fill in the missing values. In the case of multiple imputations [37,38], which we use in this study, the filling-in happens multiple times in multiple ways so that the results represent the average result across many possible ways the complete dataset might have looked. In contrast, data augmentation is using the existing data to create new data rather than fill in "holes" in the data. More generally, data imputation is estimating the results as if we knew the values of the missing data while data augmentation is creating new data to simulate a bigger dataset.

III. METHODS

In this section, we describe how we collected and processed the data, how we converted undergraduate institutions into data meaningful to our model, the algorithms we used, and how we implemented them.

A. Preparation

Data for this study come from applications to the physics graduate program at Michigan State University to enroll in fall 2014 through fall 2020. The admissions process is unique at this university in that the applications are not only reviewed by a central committee but also by members of the subdisciplines in which the applicant expresses interest. Domestic and international applicants do not undergo the same review process and hence we only analyze applications from domestic applicants. Here, a domestic applicant is defined to be a U.S. citizen or permanent resident.

Applicants submitted general and physics GRE scores, transcripts, a personal statement, a research statement, and letters of recommendation. Per a ballot initiative in the state of Michigan, Michigan State University and the other Michigan public universities are explicitly prohibited from discriminating against or granting preferential treatment to individuals based on race, sex, color, ethnicity, or national origin in education [39]. To comply with this law, our university's admissions system collects limited demographic data and our department chose not to record the information that was available when evaluating applicants.

Data from applicants planning to enroll between 2014 and 2017 were obtained through departmental spreadsheets that recorded key information from the applicants as compiled by the admissions chair. These data included

TABLE I. The three models compared in this paper and the data that went into each.

Name	Data source and features	Number of domestic applicants	Percent admitted	Where results are reported
Dataset 0	Information pulled from the applications before our department implemented a rubric (2014–2017). Features are shown in Table II.	512	48%	Section IV A
Dataset 1a	Information pulled from the *applications* after our department implemented a rubric (2018–2020). Features are shown in Table II.	511	34%	Section IV A
Dataset 1b	Rubric ratings generated *by* faculty as they evaluated applications (2018–2020). Features are shown in Table III and described in the appendix of [21].	321	43%	Section IV C

general and physics GRE scores, GPA, research subfield of interest, and undergraduate institution.

Starting with the cohort to begin our program in fall 2018, the admissions committee began using rubric-based holistic review of applicants. This process involved using a rubric to rate applicants on 18 criteria, covering academic preparation, research, noncognitive competencies or personality traits, fit with the program, and GRE scores. We also obtained these ratings as compiled by the graduate chair. More details about how rubric-based holistic review works and the rubric we used can be found in Young *et al.* [21].

In addition, we manually went through the applications for this cohort to extract the same information as was available for the cohort planning to enroll between 2014 and 2017 to form a comparative dataset. Details of the process and data handling are also described in Young *et al.* [21]. Applications from 2014 to 2017 cohort were not available to us and hence, we could not perform the same process for that cohort.

For convention, we will refer to data collected before the implementation of the rubric (fall 2014–fall 2017) as *dataset 0* following the convention of using "naught" for initial time in physics and data collected after the implementation of the rubric (fall 2018–fall 2020) as *dataset 1*, following the convention of using "1" to be mean the next time the data was collected. Furthermore, data in dataset 1 that come from the applications will be referred to as the *dataset 1a* while data that come from the faculty ratings using the rubric will be referred to as *dataset 1b* data. These are summarized in Table I.

1. Describing undergraduate institutions

Because the name of the undergraduate institution does not provide useful information to an algorithm, we created new features to describe characteristics of the institutions. To describe the overall institution, we classified each institution as public or private, whether it is a minority serving institution (MSI), the region of the country it is in (such as Northeast, Southwest, etc.), and the Barron's selectivity of the institution, which describes how selective the undergraduate program is. We assume that selectivity serves as a proxy for prestige. Classifications for the first three categories were taken from the most recent Carnegie Rankings [40] while the Barron's classification came from Barron's *Profiles of American Colleges* [41]. Because the overall reputation of the applicant's undergraduate university might not describe the physics program at that university, we also included factors related to the physics program, such as the highest physics degree offered at the university and the sizes of the undergraduate program and Ph.D. program if applicable. The sizes of the undergraduate and Ph.D. programs were determined by the median number of graduates of the program between the 2012–2013 and 2015–2016 academic years for dataset 0 and 2016–2017 through 2018–2019 for dataset 1a (i.e., the years that applicants applied to the program). The programs were then classified as small, medium-small, medium-large, or large based on which quartile they fell into. We used the Roster of Physics Departments with enrollment and degree data to collect these data [42–48]. All features used in the models for datasets 0 and 1a are shown in Table II and include the scale of measurement.

2. Justifying our choice of institutional factors

Prior work has documented university pedigree is often considered in the application process because institutional quality is assumed to be a proxy for student quality [8,49]. Here, we measure institutional quality by Barron's selectivity and public or private status, with the assumption that physics faculty view private universities as more prestigious than public universities. For example, U.S. News & World, publisher of a well-known college ranking system, has not included a public university in its top 10 in the past decade and no more than 1 public university in its top 20. We also include the region of the applicant's undergraduate university to account for the fact that the institution being studied is a public university and might therefore show a preference for students from the surrounding region.

Prior work has also found faculty exhibit a tendency to admit students like themselves, though it is more common among academics who graduated from elite institutions [8].

TABLE II. Features used in our model of datasets 0 and 1a, including their scale of measurement.

Feature	Measurement scale
Undergraduate GPA	Continuous
Verbal GRE score	Continuous
Quantitative GRE score	Continuous
Written GRE score	Continuous
Physics GRE score	Continuous
Proposed research area	Categorical
Application year	Categorical
Barron's selectivity	Categorical
Region of applicant's undergraduate institution	Categorical
Type of physics program at applicant's undergraduate institution	Categorical
Size of undergraduate physics program at applicant's undergraduate institution	Categorical
Size of doctoral physics program at applicant's undergraduate institution	Categorical
Applicant attended a minority serving institution	Binary
Public or private	Binary
Output variable: Admitted status	Binary

Therefore, it is not unreasonable to expect that faculty may prefer to admit students who followed similar paths as they did, meaning students from large, doctoral institutions might be more likely to be admitted than students from smaller institutions. Additionally, we use the sizes of the undergraduate and Ph.D. programs as proxies for the perceived prestige of the physics department, assuming a more prestigious physics department attracts more students and hence graduates more students.

B. Analysis

Here, we describe the random forest algorithm, how it develops a model of the data, and determine what features have the most impact on the predictions. We also describe a statistical test to compare the performance of different machine learning models and how we implemented these analyses.

1. The random forest algorithm

To analyze our data, we used the conditional inference forest algorithm, a variant of the random forest algorithm [28] shown to be less biased when the data include both continuous and categorical variables [50] such as those used in our model (see Table II). Random forest models in general are ensembles of individual decision trees, which use binary splits of the input features to make a prediction. The predictions are then averaged and sometimes weighted over the individual trees to obtain the overall prediction of the random forest.

While there are multiple metrics used to assess random forest and other machine learning models, two of the most common are the accuracy and the area under the curve (AUC). The accuracy is simply the proportion of correct predictions made by the model. To ensure that the accuracy is not inflated by overtraining, only a fraction of the available data is used to construct the model. The rest of the data is used to evaluate the predictive power. It is this remaining data that is used to calculate the accuracy of the model. This process of splitting data into training and testing sets is often repeated multiple times to understand the variation in the accuracy or other metric of the model through a process called *cross validation (cv)*.

The AUC is defined as the area beneath the receiver operator curve of the model, which visualizes the false-positive rate against the true-positive rate. It varies between 0.5 and 1, with values greater than 0.7 signifying an acceptable model [51]. The area describes the proportion of positive cases that are ranked above negative cases in the dataset by the model. For example, for our data, the AUC would represent the proportion of all random pairs of admitted and not-admitted applicants in which the admitted applicant is classified as admitted and the not-admitted applicant is classified as not-admitted.

In addition to making predictions, the random forest algorithm can determine the importance of each feature to the model, referred to as the feature importance. For this analysis, we use two importance measures. First, we used the AUC permutation feature importance [52] as it is claimed to be less biased than the accuracy-based permutation importance when input features differ in scale (as do our factors listed in Table II) and when the predicted variable is not split evenly between the two outcomes. In practice, our previous work suggests whether we pick the AUC-permutation importance or accuracy-based permutation importance will have minimal effects on the conclusions [53]. Under this approach, each feature is randomly permuted and then passed through the model to make a prediction. The AUC is then recorded and the difference between this value and the original AUC is computed. As permuting a feature with more predictive information should result in a worse model than permuting a feature with less predictive information, a larger difference between the original AUC and the AUC with a permuted feature suggests that this feature contains more predictive information. These differences can then be used to create a relative ordering of features.

However, if the features are correlated, it is possible that the orderings may be biased or that permutations of one feature might result in unrealistic combinations of features and hence would cause the model to extrapolate performance [54]. For example, if all students who earned perfect scores on the physics GRE also had high GPAs, permuting

GPA could cause there to be cases where a perfect physics GRE score goes with a low GPA, which would be outside of the region learned by the model. To prevent that, we use a second importance measure that has been proposed in which features are permuted within a subset of similar cases [55]. Because of the correlations between various sections of the GRE (e.g., Verostek *et al.* reported a moderate correlation between the physics GRE score and the quantitative GRE score [14]), we also used this conditional approach to compute feature importances.

Feature importances are derived from the data and hence, are not assumed to follow any statistical distribution. Therefore, there is no simple way to apply the idea of statistical significance to feature importances. We instead applied the recursive backward elimination technique described in Díaz-Uriarte and Alvarez de Andrés [56] to determine which features are predictive of admission and which are not. When using this technique, the features are ordered according to their importance. A model is then built using all the features and the accuracy is computed. A set fraction of the features with the smallest importances is then removed and a new model is built and the accuracy is computed. This process continues until only two features are left. The model with the fewest number of features while maintaining an accuracy within a standard error of the highest accuracy across all models built in this process is then the selected model. We will refer to the features used in this selected model as the *meaningful* features and interpret them as the features that are predictive of the outcome. For more information about random forest models, biases, and feature importance measures, see the supplemental material of Young *et al.* [57].

2. Comparing different classification models

When using multiple classification models on a dataset, an important consideration is how to compare the different models and determine the best one. Simple methods to do so include comparing a metric of interest such as the accuracy or the AUC and choosing the model with the highest average value over the datasets or picking the model that has the highest metric on the largest number of datasets [58].

However, it is also possible that one model may appear to be better than another due to chance. Therefore, a test of statistical significance may be of interest to better understand whether that might be the case. Dietterich [59] compared five such methods for doing so and Alpaydm [60] developed a more robust version of the 5×2 cv paired t test method preferred by Dietterich. We describe Alpaydm's 5×2 cv combined F test below.

Assume that there are two classifiers A and B and a dataset D. Split D randomly in half, forming a training set and a testing set. Then use the training set to build a model with classifiers A and B and apply those models to the testing set to obtain accuracies $p_A^{(1)}$ and $p_B^{(1)}$. Next, swap the training and testing sets and repeat the procedure, computing testing accuracies $p_A^{(2)}$ and $p_B^{(2)}$. Following this, the differences in testing accuracies between model A and B are computed, $p^{(1)}$ and $p^{(2)}$. Finally, the mean and variance of the differences are computed.

This procedure is then repeated 5 times. The f static proposed by Alpaydm is then

$$f = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}(p_i^j)^2}{2\sum_{i=1}^{5} s_i^2}, \tag{1}$$

where p_i^j is the difference in accuracies for the jth trial of the ith iteration, and s_i^2 is the estimated variance which is given by $s_i^2 = (p_i^{(1)} - \bar{p}_i)^2 + (p_i^{(2)} - \bar{p}_i)^2$ where $\bar{p}_i = (p_i^{(1)} + p_i^{(2)})/2$. This f is then approximately distributed as an F statistic with 10 and 5 degrees of freedom. The f and the degrees of freedom can then be used to calculate the probability of obtaining results given that there is no difference between classifiers A and B, the p value. If the p value is less than some cutoff, α, then the classifiers are said to be statistically different. See Alpaydm for details [60].

3. Implementation

The implementation of the analysis follows the framework detailed in Aiken *et al.* [61].

To perform the analysis, we used R [62] and the `party` package [50,55,63] to create a conditional inference forest model. We used 70% of our data to train the model, 500 trees to build our forest, and used $\sqrt{p}$ as the number of randomly selected features to use to build each tree, with p being the total number of features in the model. These values follow the recommendations of Svetnik *et al.* [64]. We ran our model 30 times, randomly selecting 70% of our data for training each time. For each trial, we calculated the training AUC, testing AUC, testing accuracy, null accuracy, and the permutation AUC importances. We then averaged the results. As the conditional inference forest algorithm has routines built in to handle missing data [65], applicants with missing information were not removed from the dataset. However, the conditional importance approach requires there to be no missing values so we used the multivariate imputation by chained equations (MICE) algorithm [66] to fill impute missing data in that case, following Nissen *et al.*'s recommendation for physics education research (PER) [67]. The imputation results were pooled using Rubin's rules [37].

For datasets 0 and 1a, the same features were used as in Table II, with the size of the physics program factors updated with new data for the postdata models. For dataset 1b, all features were treated as categorical (0, 1, or 2), and as in our previous work [21], any values between a rubric level were rounded up.

In addition, to determine if our models depended on our choice of hyperparameters, we varied the fraction of data to train the model, the number of trees in the forest, and the number of randomly selected features to use to build each tree. If we are to trust the models we created, we would expect to see minimal variation in the results based on our hyperparameters. We set the training fraction to be either 0.5, 0.6, 0.7, 0.8, or 0.9, the number of trees in the forest to be 50, 100, 500, 1000, or 5000, and the number of features used for each tree to be 1, $\sqrt{p}$, $p/3$, $p/2$, or p for a total of 125 possible combinations (124 new and the original model). These choices are based on findings in Svetnik *et al.* [64]: namely, that the error rates level off once the number of trees is on the order of 10^2 and their choices of the number of features in each tree. In addition, increasing the training fraction may improve performance as there is more data for the model to learn from. For each combination, we repeated the procedure in the previous paragraph. Due to the computational cost of the conditional permutation approach, we only calculated the AUC-permutation importance.

To determine whether changing the hyperparameters affected our models, we computed the minimum, median, and maximum values of each metric over the 125 hyperparameter combinations and the relative ordering of the features in each model. We chose the minimum, median, and maximum instead of the mean and standard error because (1) we are looking across different models rather than getting repeated measurements of the same things so we cannot assume the results will be normally distributed and (2) we are interested in the best and worst performance achieved under hyperparameter tuning to get a sense of the possible values we can achieve. This analysis would not be possible using the mean and standard error. If our model is unaffected by the choice of hyperparameters, we would expect the metrics to show minimal variation and the relative ordering of the features to be largely unchanged.

To compute the Tomek Links, we used the `TomekClassif` function in the `UBL` package [68]. We first used MICE to impute the data before calculating the Tomek Links using the function defaults with the exception of the distance metrics. Following the recommendation of the package's documentation, we used the `HVDM` distance for datasets 0 and 1a because those datasets contain both categorical and continuous data and we used the `Overlap` distance for dataset 1b because all features were categorical.

After removing the Tomek Links, we ran each model 30 times and averaged the results. Results were then pooled using Rubin's Rules.

In addition to looking at the feature order to determine if the admission process changed, we can compare the performance of the models themselves. If the process did not change, then a model built from dataset 0 should perform equally well (within error) on a dataset 0 testing set as on dataset 1, and a model built from dataset 1a should perform equally well (within error) on a dataset 1a testing set as on dataset 0. If the process did change, we would expect better performance on the test data pulled from the train and test split than the other dataset. In this approach, we are using model fit as a proxy for whether the process changed.

To test this hypothesis, we first randomly split dataset 0 into a training and testing set (70% again to the training set) and built a conditional inference forest model on the training set. We then used the model to predict the testing set and dataset 1a, computing the accuracy and AUC. We repeated this process 30 times and averaged the accuracies and AUCs. We then repeated the process for dataset 1a by doing a train and test split on dataset 1a and using all of dataset 0 as a testing set. This method provides a simple comparison between the models.

Second, we performed the 5×2 cv combined F test explained by Alpaydm [60]. Because our models were not different algorithms, we altered the approach as follows. Both dataset 0 and dataset 1a were divided into a training and test set, with half of the data in each. For each pair of trials, we used the two training datasets to develop two models (one for the before rubric process and one for after) and applied those models to the testing set from dataset 0. We then used the testing set from dataset 0 and the testing set from dataset 1a to develop two new models and applied those models to the original dataset 0 training set. The accuracies and AUCs were then subtracted for the same testing set. We then repeated this process 5 times and computed the f statistic to determine if the models were equally effective at predicting the data before the implementation of the rubric (dataset 0). To determine if the models were equally effective at predicting the data after the implementation of the rubric (dataset 1a), we repeated the procedure above, except for swapping the roles of dataset 0 and dataset 1a.

In both cases, a corrected p value less than 0.05 would signify a statistically significant difference between the predictive abilities of the models. To correct the p values for multiple comparisons, we use the Holm-Bonferroni procedure [69].

IV. RESULTS

A. Determining what drove the admissions process before and after the implementation of the rubric

To check that our model is an appropriate fit for the data, we first looked at the model's metrics. Across the 30 runs on dataset 0, the average accuracy of our model predicting the held-out data was $75.6\% \pm 0.6\%$, the average training AUC was 0.849 ± 0.002, and the average testing AUC was $0.756 \pm .006$. As our model's accuracy is significantly higher than the null accuracy of 52.7%, the percent of students who were not accepted, and our testing AUC is

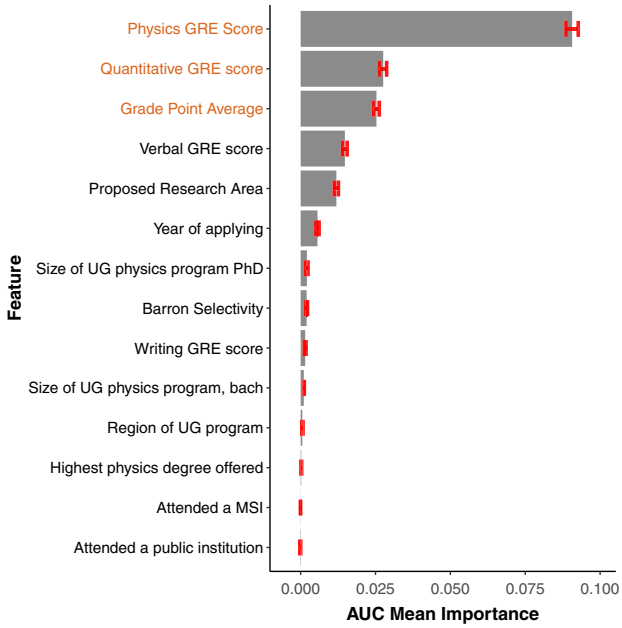

FIG. 2. Averaged AUC feature importances over 30 trials for dataset 0. Physics GRE score, Quantitative GRE score, and undergraduate GPA, appearing in orange, were the factors found to be meaningful and hence predictive of being admitted for dataset 0.

above 0.7, our model can be considered an acceptable model of the data.

The feature importances averaged over the 30 runs are shown in Fig. 2. We find numerical factors such as the applicant's score on the physics GRE, the applicant's score on the quantitative GRE, the applicant's undergraduate GPA, the applicant's verbal GRE score, and their proposed research area are more important in the application process than any factor describing the applicant's undergraduate institution. Using recursive backward elimination to determine the meaningful factors, we find the applicant's physics GRE score, quantitative GRE score, and undergraduate GPA to be the only meaningful factors.

To verify that the applicant's physics GRE score, quantitative GRE score, and undergraduate GPA were indeed the only meaningful factors before the implementation of the rubric, we then reran our random forest model 30 times using only these three factors as the predictors. Our average testing accuracy was then $75.4\% \pm 0.6\%$ and our testing average area under the curve was 0.754 ± 0.006. As we would expect when using only the meaningful features, these are not statistically different from the values we found using all 14 factors shown in Table II.

When looking at the data from after the implementation of the rubric, we find that we are less successful in building the models. Across the 30 runs on dataset 1a, the average accuracy of our model predicting on the held-out data was $71.4\% \pm 0.6\%$, the average training AUC was 0.720 ± 0.004, and the average testing AUC was $0.626 \pm .006$, which is less than the minimum of 0.7 for a reasonable model. Our null accuracy was 66.0%, meaning that our model is only doing slightly better than if it were to predict everyone was not admitted to our program as most applicants were not admitted.

Due to poor model fit, feature importances from dataset 1a should be interpreted with caution. As such, they are not included here, but for completeness, they are provided in Fig. S1 in the Supplemental Material [70] without discussion.

B. Comparing the underlying admissions models

When looking at the results, which are shown in Figs. 3 and 4, we see that models built on one dataset do not work sufficiently well on the other. In Fig. 3(a), we see that the dataset 0 test AUC is larger than the dataset 1a AUC, and in Fig. 4(a), we see that the dataset 0 test accuracy is larger than the dataset 0 null accuracy while the dataset 1a test accuracy is smaller than the dataset 1a null accuracy. These metrics suggest that the dataset 0 model fits dataset 0 well but does not fit dataset 1a well and therefore that the process might have changed.

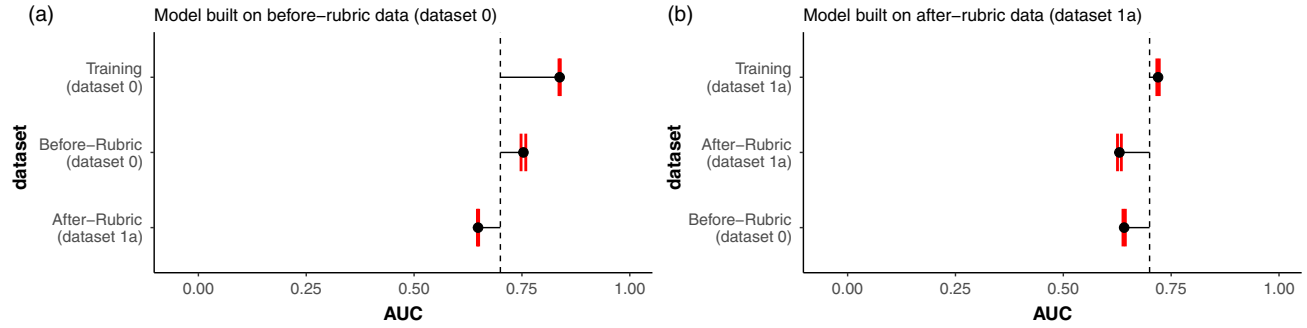

FIG. 3. Comparison of the testing AUC when (a) dataset 0 is used to train the model and (b) when dataset 1a is used to train the model. Training refers to the training AUC for the model. All error bars are 1 standard error. Results were averaged over 30 trials. When training on dataset 0, we are able to produce an acceptable AUC when testing on dataset 0 but not dataset 1a. When training on dataset 1a, we are not able to produce an acceptable AUC in either case.

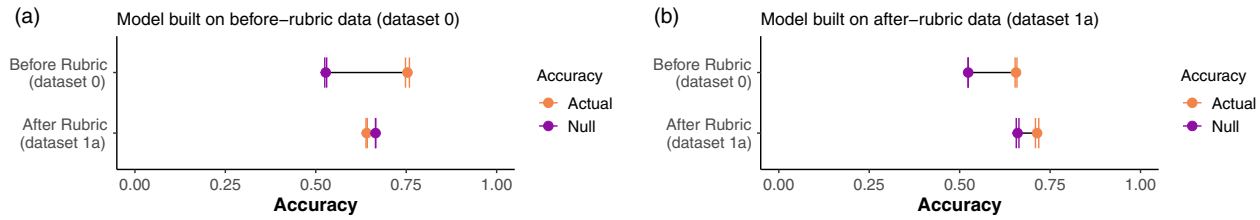

FIG. 4. Comparison of the testing accuracy when (a) dataset 0 is used to train the model and (b) when dataset 1a is used to train the model. The null accuracy is shown in cyan with the shorter in height error bars. All error bars are 1 standard error. Results were averaged over 30 trials. When training on dataset 0, we are able to produce an acceptable accuracy when testing on dataset 0 but not dataset 1a. When training on dataset 1a, we could not produce an acceptable accuracy in either case.

Looking at Fig. 3(b), we see that none of the metrics are especially good. The test AUCs are both in the poor range, suggesting that the model built from dataset 1a does not fit that well in the first place. It is then not surprising that the model does not predict dataset 0 well. Given that the initial model did not fit the data well, we cannot use the result to make a claim about whether the process changed.

When we look at the 5×2 cv combined F test, which compared a model for admission before the implementation of the rubric to a model for admission after the implementation of the rubric on the two datasets using two performance metrics, we see similar results (Table IV). We find that the models for before and after the implementation of the rubric tested on dataset 0 are statistically different while the models for before and after the implementation of the rubric tested on dataset 1a are not. However, given the results presented in Figs. 3 and 4, the lack of statistical differences for the models tested on dataset 1a is likely because both models were equally bad at fitting the data rather than a similar underlying admission process captured by the models.

TABLE III. Features used in our model of datasets 1b, including their scale of measurement.

Feature	Measurement scale
Physics coursework	Categorical
Math coursework	Categorical
Other coursework	Categorical
Academic honors	Categorical
Research: Variety or duration	Categorical
Research: Quality of work	Categorical
Research: Technical skills	Categorical
Research: Dispositions	Categorical
Achievement orientation	Categorical
Conscientiousness	Categorical
Initiative	Categorical
Perseverance	Categorical
Fit: Research	Categorical
Fit: Faculty	Categorical
Contributions to community	Categorical
Contributions to diversity	Categorical
General GRE	Categorical
Physics GRE	Categorical

C. Determining how the rubric affected what faculty emphasized in the admissions process

Given that after the implementation of the rubric applicants are rated on the rubric constructs, perhaps using the rubric constructs instead of the application data in a model would allow us to model our data better. It did not.

Looking at the model metrics to assess the fit of our model, we find that the testing AUC was 0.664 ± 0.007 and the testing accuracy was 0.675 ± 0.007 (null accuracy 0.553 ± 0.006). Given that not all applicants had sufficiently complete applications to be reviewed by faculty and those with incomplete applications tended to be not admitted, the null accuracy is smaller for models of dataset 1b than the models of dataset 1a. These results suggest that the model is not a good fit for the rubric data either.

Due to poor model fit, feature importances from dataset 1b should be interpreted with caution. As is the case for the feature importances from dataset 1a, the results without discussion are included in Fig. S1 in the Supplemental Material [70] for completeness.

D. Using Tomek Links to better understand the admissions process

Given the limited ability of the conditional inference forest to model datasets 1a and 1b, we used Tomek Links to

TABLE IV. F statistics and corrected p values for predicting on each dataset and the metric used to assess whether the predictions of the models built on datasets 0 or 1a were different. The significant p values for testing on dataset 0 suggest the models are different while the nonsignificant p values for testing on dataset 1a is likely due to poor model fit in general.

Data tested on	Metric	f	Corrected p value
Dataset 0	AUC	18.95	0.01
Dataset 0	Accuracy	9.70	0.03
Dataset 1a	AUC	1.54	0.33
Dataset 1a	Accuracy	4.14	0.13

TABLE V. Metrics when using Tomek Links and MICE for each of the three datasets. We find that using Tomek Links and MICE resulted in slight increases in the studied metrics.

	Dataset 0	Dataset 1a	Dataset 1b
Cases dropped	11%–14%	15%–18%	12%–17%
Training AUC	0.880 ± 0.004	0.760 ± 0.015	0.779 ± 0.010
Testing AUC	0.809 ± 0.009	0.670 ± 0.015	0.704 ± 0.014
Testing accuracy	0.806 ± 0.009	0.775 ± 0.012	0.717 ± 0.012
Null accuracy	0.539 ± 0.006	0.699 ± 0.009	0.575 ± 0.010

remove boundary cases. As the goal was to build models that better fit the data and hence, whose outcomes we could place more trust, we focused on the model metrics instead of importances. That was because the importances remained more or less unchanged. The results are shown in Table V. As MICE generates new values for each imputation and hence, affects which cases are nearest neighbors, the percent of cases dropped for each trial varies.

First, we notice that for dataset 0, using Tomek Links increased the testing AUC and testing accuracy by 0.05 over the original model. The testing AUC is now about 0.8 which is considered "good" compared to "fair" for the original model [51].

Likewise, using Tomek Links also results in an approximately 0.05 increase in the testing AUC and testing accuracy for dataset 1a. However, the AUC is still in the poor range and the testing accuracy is only slightly better than the null accuracy.

For dataset 1b, using Tomek Links increases the testing AUC and testing accuracy by approximately 0.04. This time, the increase in the testing AUC is enough for the model to be classified as "fair."

To better understand what Tomek Links were doing in the modeling process, we investigated how removing the boundary cases affected the decision boundary. To plot the results, we only used the physics GRE score and undergraduate GPA to make a simple model for datasets 0 and 1a. To compute the Tomek Links, we used MICE to create a complete dataset first and then found the Tomek Links. As all the data in dataset 1b was categorical, a 2D plot of the decision boundary would have yielded limited insight and hence, we did not do so. The results of a single trial are shown in the Supplemental Material [70].

For both cases, we find that using Tomek Links appears to reduce the overfitting. Applicants with higher physics GRE scores and higher GPAs were predicted to be admitted while applicants with lower physics GRE scores and GPAs were predicted not to be admitted.

For the feature importances, we find that the ordering of the features is more or less the same as presented in Figs. 2 and S1 in the Supplemental Material [70]. As the results are not too different, the plots are relegated to the Supplemental Material [70].

V. VALIDATING OUR RESULTS

To increase our trust in our results and show that our results are not artifacts of how we modeled the data, we reanalyzed our data using alternative model specifications and taking correlations into account. Ideally, these would not change our results from dataset 0 where we were able to create an acceptable model but would improve our results from datasets 1a and 1b where we were unable to create acceptable models. Such a result in the latter case would mean that our initial models were not considering all relevant effects.

A. Dataset 0

When we test the various hyperparameter combinations for the data before the implementation of the rubric (dataset 0), we find similar results as we did originally. Looking at the metrics (Table VI), we see that the testing accuracy varies by 3.3 percentage points between the minimum and maximum values and the testing AUC varies by 0.034 between the minimum and maximum values. As the variation is limited and these metrics are still within the acceptable range, the results suggest that our choice of hyperparameters has limited impact on the metrics.

When we look at the ranks of the features used in each hyperparameter combination, we also see limited variation. For interested readers, the relevant plots are included in the Supplemental Material [70]. First, we find that physics GRE score, GPA, quantitative and verbal GRE scores, and proposed research area are always the top five features for models built on the data before the implementation of the rubric, regardless of the hyperparameters. Second, we find that the institutional features never rank in the upper half of the features, meaning that no combination of hyperparameters can create a model where these features are predictive of admission. In addition, we notice that the year of applying is always ranked sixth, serving as a separating feature from the previous two groups of features. This result is likely because there are yearly differences in the fraction of applicants admitted so year is not a noise feature and should be ranked above the noise features. However, knowing the year the applicant applied does not say too much about the applicant themselves and hence, we

TABLE VI. Minimum, median, and maximum values of the metrics obtained over the 125 hyperparameter combinations for models built from dataset 0. The results suggest that hyperparameter tuning results in minimal changes in the metrics.

Metric	Minimum	Median	Maximum
Train AUC	0.824	0.848	0.853
Test AUC	0.726	0.749	0.760
Test accuracy	0.727	0.750	0.760
Null accuracy	0.521	0.527	0.556

would expect it to rank below the features like test scores and GPA that describe the applicant.

Looking at the most important features for dataset 0 more closely, we notice that physics GRE is always the top-ranked feature followed by either GPA or quantitative GRE score, with GPA being the more common selection. Furthermore, GPA never ranks lower than third while the quantitative GRE score ranks between second and fourth. For certain choices of hyperparameters, the applicant's proposed area of research ranks higher than the quantitative GRE score.

We also find limited differences in the results when using the conditional feature importances on dataset 0, the exception being that the quantitative GRE score is no longer meaningful. Those results are also shown in the Supplemental Material [70].

B. Dataset 1a

Given the poor performance of our model on dataset 1a (AUC < 0.7, testing accuracy only slightly higher than the null accuracy), hyperparameter tuning might have improved the model. While it did to a degree, the testing accuracy was still only a few percentage points above the null accuracy and the testing AUC was still below 0.7 (Table VII). Thus, even with hyperparameter tuning, the models of dataset 1a were poor.

As the model fits were still poor, we do not interpret the resulting feature importances. However, the occurrence fraction of each rank for each feature and the conditional importances are again included in the Supplemental Material [70].

C. Dataset 1b

As was the case for dataset 1a, we were unable to make a substantially better model for dataset 1b. Even the best AUC among the 125 hyperparameter tuning combinations did not exceed 0.7. The full results are shown in Table VIII.

As is the case for dataset 1a, because the model fits were still poor, we do not interpret the resulting feature importances. The occurrence fraction of each rank for each feature and the conditional importances are again included in the Supplemental Material [70].

TABLE VII. Minimum, median, and maximum values of the metrics obtained over the 125 hyperparameter combinations for models built from dataset 1a. The results suggest that hyperparameter tuning results in minimal changes in the metrics.

Metric	Minimum	Median	Maximum
Train AUC	0.602	0.735	0.749
Test AUC	0.549	0.633	0.676
Test accuracy	0.679	0.712	0.732
Null accuracy	0.645	0.661	0.666

TABLE VIII. Minimum, median, and maximum values of the metrics obtained over the 125 hyperparameter combinations for the models of dataset 1b. The results suggest that hyperparameter tuning results in minimal changes in the metrics.

Metric	Minimum	Median	Maximum
Train AUC	0.711	0.767	0.791
Test AUC	0.654	0.669	0.686
Test accuracy	0.660	0.678	0.696
Null accuracy	0.559	0.561	0.586

VI. DISCUSSION

Here, we first provide answers to our research questions and then use those answers to address the larger question of whether our department's admissions process changed.

A. Research questions

1. What parts of a graduate application determine whether an applicant will be admitted to a physics program?

For dataset 0 and our process before the implementation of the rubric, we found the applicant's physics GRE score, quantitative GRE score, and GPA to be predictive of admission. The general result of quantitative metrics being most important to the admissions process aligns with previous work that examined the process from the perspectives of faculty [5,8].

For dataset 1a and our admission process after the implementation of the rubric, we were unable to build an acceptable model of the data and hence do not provide interpretation of the features predictive of admission. Conditional inference forests will always return importance values regardless of how well the model fits and therefore we should interpret the results with a degree of caution when the model does not fit the data well. Even after hyperparameter tuning, we were unable to achieve a testing accuracy of more than a few percentage points above the null accuracy or a testing AUC above 0.7, suggesting a poor model for dataset 1a. In contrast, we were able to successfully model dataset 0 and feel more confident that we are modeling the underlying admissions process rather than random variations in the data.

Despite prior work suggesting institutional characteristics play an important role in graduate admissions, we did not find institutional or departmental characteristics to be meaningful to models of dataset 0 and hence, predictive of admission. Our result could be due to differences in methodology or due to institutional effects being influential but not dominant factors [29]. Indeed, Posselt suggests institutional factors might be used to differentiate applicants with similar GPAs and GRE scores [8]. Therefore, we might not have found institutional factors to be predictive of admission because they are used when primary factors

such as GPA and GRE scores do not sufficiently separate applicants.

We also note that any data about the written components of the application (e.g., personal statements, research statements, letters of recommendation, etc.) are not present in these datasets so we cannot make any conclusions about how those may change the previous orders or the results. Our results should then be interpreted in the context that the quantitative measures are assumed to be the sole reasons for applicants being admitted. In practice, quantitative measures are not the sole reason applicants are admitted but previous qualitative studies have found that they are at or near the top of the list (e.g., [5,8]). These studies then suggest that our assumption is not entirely unrealistic.

2. How does the introduction of a rubric with defined constructs to evaluate applicants affect which parts of the application determine admission?

Because we were unable to build acceptable models of the admissions process following the introduction of the rubric, we cannot determine how the introduction of our rubric affected which parts of the application determine admission. However, we were able to create acceptable models of the data before the implementation of the rubric and identify the parts of the application that determine admission but not after suggesting that the parts of the application that drive admissions decisions have changed.

When modeling the data after the implementation of the rubric, we note that using the rubric features does result in improved metrics compared to the traditional features for the data collected after the implementation of the rubric. However, the metrics are still outside of the acceptable range for trusting our models are capturing the underlying process. One possible reason for that result is that dataset 1b has a less imbalanced outcome.

To see if that was the case, we created a model using the data in dataset 1a that corresponded to the applicants in dataset 1b. When we did so, we found that the metrics were comparable, but the original test dataset 1b model slightly outperformed this new model (0.02 increase in testing AUC and accuracy). Thus, while some of the improvements in metrics might be attributable to the more balanced dataset, using the rubric constructs also provided some benefits. However, it was still not enough for us to trust the results of the models on dataset 1b.

3. How does using Tomek Links affect our ability to answer the first two research questions?

Despite the hope that removing potentially problematic cases from the data via Tomek Links would provide additional insight into the first two research questions, which did not manifest in practice.

We did not find much difference in the parts of the application that mattered most for admission to our graduate program. However, the results do provide evidence that the underlying processes for admission might have changed. For dataset 1b, using Tomek Links increased the testing AUC over 0.7, which is considered "fair." However, while using Tomek Links for dataset 1a did improve the testing AUC, it did not do so enough for the model to be considered acceptable. These results provide evidence that faculty were using the rubric to make admissions decisions rather than continuing the original process. However, claiming that the process changed solely off the fact that we can produce an AUC over 0.7 for dataset 1b is unwarranted.

More generally, while the benefits were relatively small, these results suggest that Tomek Links are a promising technique for modeling PER data and should be investigated further. They can be especially useful for datasets where we expect many boundary cases or cases that go against the general trend. For example, if we were to predict who passes an introductory class, Tomek Links might allow us to remove students who earned exam scores around the minimum passing grade and thus might or might not have passed the course or anomalous students who did poorly on the midterms but managed to earn a high grade on the final to pass the class.

B. Addressing whether our process changed

Looking across the research questions, we can now address our larger question of *did the introduction of the rubric change our department's admissions process*. Overall, the evidence points in the direction of the process changing.

In terms of evidence for the process changing, we find that the models of datasets 1a and 1b do not fit the data well. As we were able to fit the dataset 0 models to an acceptable degree using the conditional inference forest algorithm but not the models of datasets 1a or 1b, this result seems to imply that there must be something different about the datasets. Because dataset 0 and dataset 1a used the same features, it is hard to explain why we could model one well but not the other unless the "true" models of the data were different and hence the admission process changed.

In addition, a model trained on dataset 0 was better able to predict held-out data from dataset 0 compared to dataset 1a. In addition, the 5×2 cv combined F test found statistically significant differences in the performance of the models. If the process had not changed, we would have expected the predictive performance to be similar.

Finally, using Tomek Links to remove applicants who might have gone against the general trend resulted in minimal increases in the metrics for the models of datasets 1a and 1b. If the process did not change, we would expect that removing applicants who might have gone against the overall trend would have led to a better model because we were able to model the admissions data before the implementation of the rubric. Yet, that is not what happened,

suggesting again there must be something different about the data collected after the implementation of the rubric.

C. Limitations affecting our ability to address whether the process changed

Looking at the results, it is possible that someone could instead believe the results suggest the process did not change. We address those here.

In terms of evidence for the process not changing, our results show that the most predictive features are similar regardless of which dataset we used. When using dataset 0, we found that the physics GRE, quantitative GRE, and GPA were most predictive of admission. Likewise, when looking at dataset 1a, we found that the physics GRE, GPA, quantitative GRE, verbal GRE, and proposed area of research were the most predictive. Using dataset 1b showed the most differences in that the measures of grades and the general GRE scores were in the lower half of the rankings. However, the physics GRE was still the top-ranked feature. Yet, both models of the data after the implementation of the rubric did not have acceptable testing metrics, suggesting that we should interpret the feature importance orders with caution. Conditional inference forest models will always produce feature importances regardless of how well the model fits the data. Because the metrics to assess fit are relatively poor, we should not trust the conclusion that the most predictive features are the same between these models.

However, it is possible that the low metrics might be a result of the conditional inference forest method not being suited to the data we have. Recent work suggests that the conditional inference forest algorithm does not perform well with missing data [71]. When we used MICE to impute the missing data, the models were still not able to produce testing metrics in the acceptable range, suggesting that the missing data were not the issue. In addition, a recent study using admissions data to predict later performance in a graduate program found that random forest methods were among the best-performing methods compared to other common methods such as logistic regression, support vector machines, Naive Bayes, and neural networks, suggesting that our choice of algorithm is unlikely to be creating the observed poor performance [72].

In addition, while conditional inference forests were designed to better handle categorical data than traditional random forests do, there could still be issues with categorical data. For example, for dataset 1b, there are only three possible values for each feature. Therefore, the model can only split each feature in three ways, which limits the depth of the trees and the fine-tuning of the model. However, when we used the section total (which could take on any integer between 0 and 8), the results did not substantially improve, suggesting that the scale of the data may not be to blame.

Even if the number of categories does not matter, the fact that some of the categorical data are discretized, continuous features (e.g., physics GRE score, physics coursework) could create problems. Prior work has shown that binning continuous features can lead to a loss of information and overestimation or underestimation of effect sizes [73,74]. It is possible that such an effect is present in our data. However, models built from datasets 1a and 1b both found the physics GRE score to be the top feature even though the physics GRE score was discretized in dataset 1b. Because the model metrics were not great (the testing accuracy was only a few percentage points above the null accuracy and the testing AUC was less than 0.7), this rebuttal should be treated with caution. On the other hand, the fact that models of dataset 1a, where discretization was not an issue, still had poor metrics suggests that it cannot fully explain the models' low metrics.

It is also possible that the low metrics are not a result of how we handled the data we had, but rather what data we had. It is possible that the applicant pools differed substantially before and after the implementation of the rubric or that committee members were using something not included in our data to evaluate applicants and if we had that data, our models of dataset 1a and 1b would improve. An analysis of the applicant pools (included in the Appendix), suggests the applicant pools are not substantially different on key measures and while such an explanation about extraneous features seems possible for dataset 1a, it seems unlikely for dataset 1b because members of the department decided what qualities they wanted to evaluate applicants on and added them to the rubric.

In addition, it is possible that datasets we had were too small for us to properly model. That is, if datasets 1a and 1b were larger, perhaps we would have been able to produce models with acceptable testing metrics and hence, trust the importance rank results. However, given that dataset 0 and dataset 1a were of similar sizes, it would be difficult to explain why we were able to create acceptable models for dataset 0 but not dataset 1a if the underlying admission processes were the same.

Finally, it is possible that the low metrics might not be caused by the data or the model and instead, the low metrics could be caused by the admission process itself. The goal of the rubric is to rate applicants along multiple dimensions, and hence in a holistic manner. If applicants were actually assessed holistically, we would expect that the model would not generalize well because there is no single underlying process. Instead, there might be multiple routes an applicant could take to gain admission and hence, the model might encounter difficulties modeling this process. The fact that hyperparameter tuning and Tomek Links did not increase the testing metrics to an acceptable range for models of dataset 1a and barely did so for the models of dataset 1b supports such an interpretation. However,

claiming the process is more holistic based on these results alone is premature, especially given the relatively small number of applicants in dataset 1b. Instead, results from other modeling attempts would either need to show poor predictive ability or show evidence of multiple routes to admission to support such a claim.

VII. FUTURE WORK

To better address the limitations of our study and to consider whether rubric-based holistic review is actually holistic in nature, future work should examine alternative techniques for analyzing this and similar admissions processes.

First, to determine if our current process is actually holistic as opposed to different from our original admissions process, future work could analyze the data using cluster analysis or latent class analysis. While such methods are becoming popular for analyzing learning environments (e.g., see [75,76]), to our knowledge, such methods are less common in studies of graduate admissions processes. To our knowledge, clusteringlike techniques have only been used to understand admissions strategies based on surveys of faculty on admissions committees [10]. If the process is more holistic, such methods might be able to identify clusters of applicants who were admitted for similar reasons. For example, some applicants may be admitted due to stellar academic credentials, others may be admitted due to their research background, while others may be admitted based on which faculty members are seeking new students. Finding or not finding such a result would provide greater clarity as to how the process may have changed. To do so, however, would require a larger dataset, especially if there are a large number of driving results for why an applicant is admitted.

Second, future work could take a mixed methods approach by considering qualitative approaches to investigate how our admissions process might have changed. Such qualitative approaches could allow us to observe the admissions process itself (similar to the studies Posselt conducted as documented in [8]) and understand how faculty are evaluating and discussing applicants in real time. In addition, a qualitative approach would allow us to avoid many of the modeling limitations related to the scale of the data and metrics while also gaining a richer understanding of similar applicants with differing admissions outcomes.

Finally, future work could directly ask faculty who have served on the admissions committee both before and after the implementation of the rubric about their perception of the process at each time. However, we must be careful of faculty's potential biases when recalling how things were done in the past (see Müggenburg for an overview [77]). For example, given the greater emphasis on diversity and equity in higher education now, faculty's recall may suffer from postrationalization [78] where they justify their decisions using reasons that were not available at the time but are consistent with their current self-image or social desirability [79] where past events may be distorted to conform to current attitudes and norms. Such an approach would be better aligned for departments considering but have not yet switched to a rubric-based holistic review. In that case, faculty could be interviewed before and after changing the admissions process.

More broadly, future work should consider the admissions process at other physics departments and understand how changes designed to make the process more equitable work in practice at other institutions. This study was done at a primarily white institution (PWI) and might not be applicable to universities with differing applicant populations. While Kanim and Cid note that having a relatively homogeneous research sample can be valuable for reducing variability, especially in early studies, they also note that exploring the effects of variability can lead to new results and a greater understanding of the results [80]. Thus, while our results might generalize to many physics graduate programs, it might also hide important differences in features predictive of admission for applicants of different demographic groups and institutions with different demographics than our own.

VIII. CONCLUSION

Overall, the results of this initial investigation are suggestive that our admission process did change after the implementation of the rubric. We were able to model the data before the implementation of the rubric to a sufficient degree but were not able to model the data after the implementation of the rubric. In addition, the model of the admissions process before the implementation of the rubric does not do well predicting the data collected after the implementation of the rubric and vice versa, suggesting that the underlying process did change. In that case, the physics GRE, GPA, and quantitative GRE seem to hold less weight in our admissions process. However, there are still numerous limitations that need to be addressed before we can make a definitive conclusion.

Furthermore, the models of the data following the implementation of the rubric performing poorly suggest that the process might be holistic. To make such a conclusion, however, we would need either evidence in favor of the occurrence of holistic admissions or stronger evidence that the current admission process is not easily modeled by known techniques. Such evidence could be obtained through a variety of quantitative or qualitative approaches.

In terms of the modeling approaches, Tomek Links seem like a promising technique for future PER studies. While their use was not enough to provide a more conclusive answer to the question of whether our admission process changed, their use did provide evidence that the data collected after the implementation of the rubric may be

modelable to an acceptable level, leaving open the possibility that other methods may be able to model the data and hence should be explored.

Finally, to truly get a sense of whether admission processes change after the implementation of a rubric or merely use a new tool to do the same process, studies such as these need to be completed in other physics departments. By doing so, we will have a better idea of how rubric-based admissions might change admission processes and whether they achieve their goal of more equitable admissions.

ACKNOWLEDGMENTS

We would like to thank Scott Pratt, Remco Zegers, and Kirsten Tollefson for providing the data for this project. We would also like to thank Tabitha Hudson for compiling the data. This project was supported by the Michigan State University College of Natural Sciences and the Lappan-Phillips Foundation.

APPENDIX: COMPARISON OF DATASETS

An alternative explanation as to why we were able to model the data before the implementation of the rubric (dataset 0) but not the data after the implementation of the rubric (dataset 1a) could be the underlying data, rather than the admissions process, is different. Here, we provide evidence to suggest that is not the case.

Given the results of Sec. IV A where we could model the data before the implementation of the rubric data, we compared the distributions of the top features from those models (dataset 0) to the distributions of those features from the data after the implementation of the rubric (dataset 1a). If the distributions of the features were statistically the same for the two datasets, it is would be difficult to explain why we could model those distributions for dataset 0, but not dataset 1a if the underlying process were the same.

The raincloud plots [81] of the distributions of the applicant's physics GRE scores, GPA, and quantitative GRE scores, the most predictive features of dataset 0, are shown in Figs. 5, 6, and 7.

From Fig. 5, we notice that the distributions of the physics GRE scores of all applicants before and after the implementation of the rubric seem similar. However, the applicants after the implementation of the rubric seem to have a slightly higher median physics GRE score. The admitted applicants after the implementation of the rubric also seem to have a similar median physics GRE score as the admitted applicants before the implementation of the rubric. In contrast, the nonadmitted applicants after the implementation of the rubric had a higher median physics GRE score than the nonadmitted applicants before the implementation of the rubric.

In Fig. 6, we see a similar result when comparing the grade-point averages of applicants before and after the

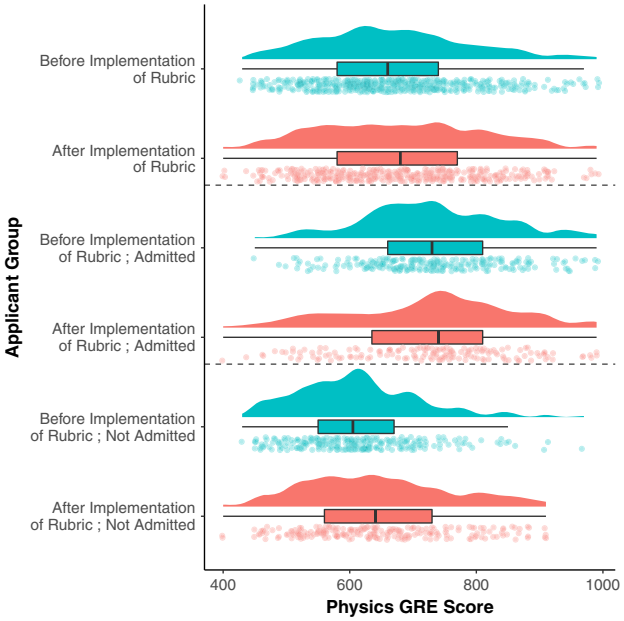

FIG. 5. Raincloud plots showing the distribution of physics GRE scores of all applicants before and after the implementation of the rubric, only admitted applicants, and only nonadmitted applicants. Only the distributions of physics GRE scores for nonadmitted applicants before and after the implementation of the rubric were found to be statistically different.

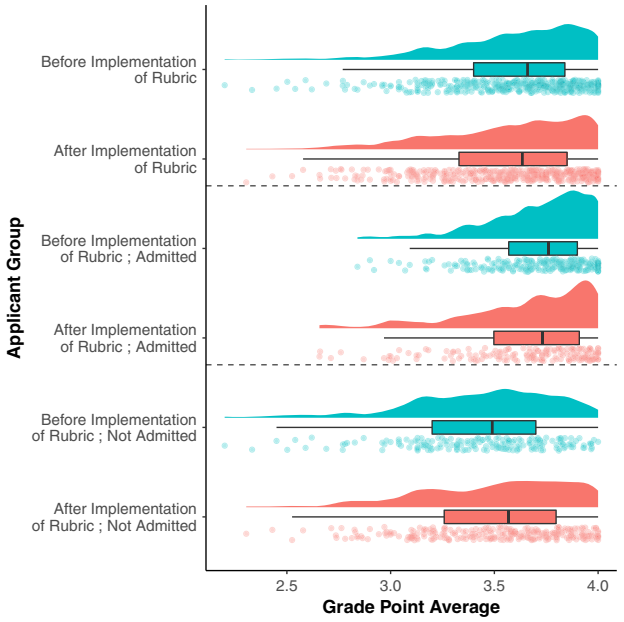

FIG. 6. Raincloud plots showing the distribution of grade-point averages of all applicants before and after the implementation of the rubric, only admitted applicants, and only nonadmitted applicants. None of the distributions of GPAs for applicants before and after the implementation of the rubric were found to be statistically different.

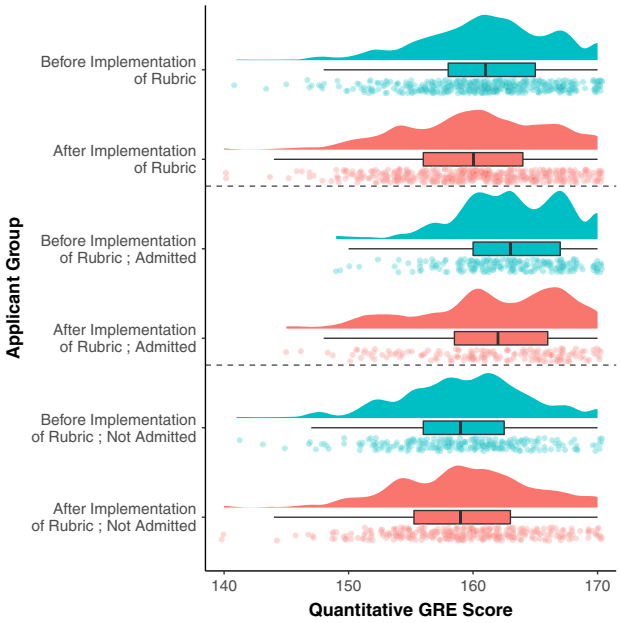

FIG. 7. Raincloud plots showing the distribution of quantitative GRE scores of all applicants before and after the implementation of the rubric, only admitted applicants, and only nonadmitted applicants. None of the distributions of quantitative GRE scores for applicants before and after the implementation of the rubric were found to be statistically different.

implementation of the rubric as well as when we break applicants into admits and nonadmits.

In Fig. 7, we see applicants after the implementation of the rubric had a lower median quantitative GRE score than applicants before the implementation of the rubric. The same is true for admitted applicants while nonadmitted applicants had similar median quantitative GRE scores, regardless of whether they applied before or after the implementation of the rubric. To determine if these differences were statistically significant, we conducted Kolmogorov-Smirnov tests between the applicants who applied before and after the implementation of the rubric [82]. As there were nine tests (physics GRE score, GPA, and quantitative GRE score for all, admitted, and nonadmitted applicants), we used the Holm-Bonferroni method to correct p values for multiple comparisons [69]. With this method, the smallest p value is compared to $0.05/n$, the next smallest p value to $0.05/(n-1)$ and so on until the null hypothesis is not rejected. At that point, we are unable to reject any remaining null hypotheses. For this procedure, n is the total number of hypotheses tested and for this analysis $n = 9$.

TABLE IX. D and uncorrected p value from Kolmogorov-Smirnov test on distributions of applicants before and after the implementation of the rubric.

Feature	Group	D	Uncorrected p value	Significant?
Physics GRE	All	0.080	0.130	No
	Admitted	0.101	0.286	No
	Nonadmitted	0.173	0.002	Yes
GPA	All	0.064	0.371	No
	Admitted	0.091	0.404	No
	Nonadmitted	0.118	0.109	No
Quantitative GRE	All	0.091	0.032	No
	Admitted	0.128	0.077	No
	Nonadmitted	0.037	0.989	No

The results of the Kolmogorov-Smirnov tests are shown in Table IX. We find that the distributions of physics GRE scores are statistically different for nonadmitted applicants before the implementation of the rubric and nonadmitted applicants after the implementation of the rubric. For all other comparisons, we are unable to reject the null hypothesis that the distributions are the same.

Given that two of the three top features for predicting which applicants would be admitted before the implementation of the rubric were not found to have different distributions for any of the groups and the third was only found to have a differing distribution for one of the three groups, it seems that the data are not the reasons for our inability to model dataset 1a.

To further check is this claim, we reran the model on dataset 0 without using applicant's physics GRE score but all of the other features listed in Table II. When we did so, we found a testing accuracy of 0.722 ± 0.005 and a testing AUC of 0.722 ± 0.005, suggesting a decent model still. Therefore, even though the distribution of physics GRE scores for nonadmitted applicants before and after the implementation of the rubric are different, that we are still able to model the dataset 0 well enough without the physics GRE scores included suggests that the differences in distributions should not affect our ability to produce a decent enough model of dataset 1a.

[1] National Center for Science and Engineering Statistics (NCSES), Survey of doctorate recipients, National Science Foundation, Alexandria, VA, Technical Report No. NSF 21-320, 2021.

[2] C. W. Miller, B. M. Zwickl, J. R. Posselt, R. T. Silvestrini, and T. Hodapp, Typical physics Ph.D. admissions criteria limit access to underrepresented groups but fail to predict doctoral completion, Sci. Adv. **5**, eaat7550 (2019).

[3] R. Sowell, Ting Zhang, and Kenneth Redd, *Ph. D. Completion and Attrition: Analysis of Baseline Program Data from the Ph. D. Completion Project*, edited by M. F. King (Council of a Graduate School, Washington DC, 2008).

[4] D. Denecke and J. Slimowitz, *Ph. D. Completion and Attrition: Policy, Numbers, Leadership, and Next Steps* (Council of Graduate Schools, Washington DC, 2004).

[5] G. Potvin, D. Chari, and T. Hodapp, Investigating approaches to diversity in a national survey of physics doctoral degree programs: The graduate admissions landscape, Phys. Rev. Phys. Educ. Res. **13**, 020142 (2017).

[6] D. Chari and G. Potvin, Admissions practices in terminal master's degree-granting physics departments: A comparative analysis, Phys. Rev. Phys. Educ. Res. **15**, 010104 (2019).

[7] D. Chari and G. Potvin, Understanding the importance of graduate admissions criteria according to prospective graduate students, Phys. Rev. Phys. Educ. Res. **15**, 023101 (2019).

[8] J. R. Posselt, *Inside Graduate Admissions* (Harvard University Press, Cambridge, MA, 2016).

[9] J. Posselt, T. Hernandez, G. Cochran, and C. Miller, Metrics first, diversity later? Making the short list and getting admitted to physics PhD programs, J. Women Minorities Sci. Eng. **25**, 283 (2019).

[10] J. Doyle and G. Potvin, In search of distinct graduate admission strategies in physics: An exploratory study using topological data analysis, presented at the PER Conf. 2015, College Park, MD, 10.1119/perc.2015.pr.022.

[11] K. M. Whitcomb, S. Cwik, and C. Singh, Not all disadvantages are equal: Racial/Ethnic minority students have largest disadvantage among demographic groups in both STEM and non-STEM GPA, AERA Open **7**, 1 (2021).

[12] C. Miller and K. Stassun, A test that fails, Nature (London) **510**, 303 (2014).

[13] N. J. Mikkelsen, N. T. Young, and M. D. Caballero, Investigating institutional influence on graduate program admissions by modeling physics Graduate Record Examination cutoff scores, Phys. Rev. Phys. Educ. Res. **17**, 010109 (2021).

[14] M. Verostek, C. W. Miller, and B. Zwickl, Analyzing admissions metrics as predictors of graduate GPA and whether graduate GPA mediates Ph.D. completion, Phys. Rev. Phys. Educ. Res. **17**, 020115 (2021).

[15] T. Wilkerson, The relationship between admission credentials and the success of students admitted to a physics doctoral program, Doctor thesis, University of Central Florida, 2007.

[16] G. L. Cochran, T. Hodapp, and E. E. A. Brown, Identifying barriers to ethnic/racial minority students' participation in graduate physics, presented at PER Conf. 2017, Cincinnati, OH, 10.1119/perc.2017.pr.018.

[17] L. M. Owens, B. M. Zwickl, S. V. Franklin, and C. W. Miller, Physics GRE requirements create uneven playing field for graduate applicants, presented at PER Conf. 2020, virtual conference, 10.1119/perc.2020.pr.Owens.

[18] A. L. Rudolph, K. Holley-Bockelmann, and J. Posselt, PhD bridge programmes as engines for access, diversity and inclusion, Nat Astron. **3**, 1080 (2019).

[19] J. D. Kent and M. T. McCarthy, *Holistic Review in Graduate Admissions: A Report from the Council of Graduate Schools* (Council of Graduate Students, Washington, DC, 2016).

[20] C. Miller and J. Posselt, Equitable admissions in the time of COVID-19, Physics **13**, 199 (2020).

[21] N. T. Young, K. Tollefson, R. G. Zegers, and M. D. Caballero, Rubric-based holistic review: A promising route to equitable graduate admissions in physics, Phys. Rev. Phys. Educ. Res. **18**, 020140 (2022).

[22] L. Stiner-Jones and W. Windl, Work in progress: Aligning what we want with what we seek: Increasing comprehensive review in the graduate admissions process, *Paper presented at 2019 ASEE Annual Conference & Exposition, Tampa, Florida* (2020), 10.18260/1-2–33592.

[23] S. Barker and A. Clobes, Work in progress: A holistic PhD admissions rubric–design & implementation, in *Proceedings of 2021 ASEE Virtual Annual Conference Content Access* (ASEE Conferences, virtual conference, 2021).

[24] N. T. Young and M. D. Caballero, Using machine learning to understand physics graduate school admissions, in presented at PER Conf. 2019, Provo, UT, 10.1119/perc.2019.pr.Young.

[25] Ivan Tomek, Two modifications of CNN, IEEE Trans. Syst. Man Cybern. **SMC-6**, 769 (1976).

[26] J. R. Posselt, Disciplinary logics in doctoral admissions: Understanding patterns of faculty evaluation, J. Higher Educ. **86**, 807 (2015).

[27] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer Science & Business Media, 2009).

[28] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).

[29] G. Attiyeh and R. Attiyeh, Testing for bias in graduate school admissions, J. Hum. Resources; Madison **32**, 524 (1997).

[30] N. E. Barceló, S. Shadravan, C. R. Wells, N. Goodsmith, B. Tarrant, T. Shaddox, Y. Yang, E. Bath, and K. DeBonis, Reimagining merit and representation: Promoting equity and reducing bias in GME through holistic review, Acad Psych. **45**, 34 (2021).

[31] A. R. Small, arXiv:1709.02895.

[32] T. R. Hoens and N. V. Chawla, Range restriction, admissions criteria, and correlation studies of standardized tests, *Imbalanced Learning* (John Wiley & Sons, Ltd, New York, 2013), pp. 43–59.

[33] Min Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data, in *Proceedings of 2016 IEEE International Conference of Online Analysis and Computing Science*

(ICOACS), Chongqing, China (IEEE, New York, 2016), pp. 225–228, 10.1109/ICOACS.2016.7563084.

[34] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explor. Newsl. **6**, 20 (2004).

[35] S. Sawangarreerak and P. Thanathamathee, Random forest with sampling techniques for handling imbalanced prediction of university student depression, Information **11**, 519 (2020).

[36] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, Data augmentation for discrimination prevention and bias disambiguation, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20 (Association for Computing Machinery, New York, NY, 2020), pp. 358–364, 10.1145/3375627.3375865.

[37] D. Rubin, Multiple imputation for nonresponse in surveys, Survey Methodol. **12**, 37 (1986).

[38] S. van Buuren, *Flexible Imputation of Missing Data*, 2nd ed. (Chapman & Hall/CRC Press, New York, 2018), ISBN 9780429492259, 10.1201/9780429492259.

[39] Constitution of the State of Michigan—Article I, Affirmative Action Programs, Sec. 26.

[40] Indiana University Center for Postsecondary Research, Carnegie Classifications 2021 public data file (2021), http://carnegieclassifications.acenet.edu/downloads/CCIHE2021-PublicDataFile.xlsx.

[41] M. Carl, NCES-Barron's Admissions Competitiveness Index Data Files: 1972, 1982, 1992, 2004, 2008, 2014 (National Center for Education Statistics, U.S. Department of Education, 2017), https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2016332.

[42] S. Nicholson and P. J. Mulvey, Roster of Physics Departments with Enrollment and Degree Data, 2013, American Institute of Physics, Technical Report, 2014.

[43] S. Nicholson and P. J. Mulvey, Roster of Physics Departments with Enrollment and Degree Data, 2014, American Institute of Physics, Technical Report, 2015.

[44] S. Nicholson and P. J. Mulvey, Roster of Physics Departments with Enrollment and Degree Data, American Institute of Physics, Technical Report, 2016.

[45] S. Nicholson and P. J. Mulvey, Roster of Physics Departments with Enrollment and Degree Data, 2016, American Institute of Physics, Technical Report, 2017.

[46] S. Nicholson and P. J. Mulvey, Roster of Physics Departments with Enrollment and Degree Data, 2017, American Institute of Physics, Technical Report, 2018.

[47] S. Nicholson and P. J. Mulvey, Roster of Physics Departments with Enrollment and Degree Data, 2018, American Institute of Physics, Technical Report, 2019.

[48] S. Nicholson and P. J. Mulvey, Roster of Physics Departments with Enrollment and Degree Data, 2019, American Institute of Physics, Technical Report, 2020.

[49] P. Paxton and K. A. Bollen, Perceived quality and methodology in graduate department ratings: Sociology, political science, and economics, Sociol. Educ. **76**, 71 (2003).

[50] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution, BMC Bioinf. **8**, 25 (2007).

[51] M. B. Araújo, R. G. Pearson, W. Thuiller, and M. Erhard, Validation of species–climate impact models under climate change, Global Change Biol. **11**, 1504 (2005).

[52] S. Janitza, C. Strobl, and A.-L. Boulesteix, An AUC-based permutation variable importance measure for random forests, BMC Bioinf. **14**, 119 (2013).

[53] N. T. Young and M. D. Caballero, Predictive and explanatory models might miss informative features in educational data, J. Educ. Data Mining **13**, 31 (2021).

[54] G. Hooker, L. Mentch, and S. Zhou, Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance, Stat. Comput. **31**, 82 (2021).

[55] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, Conditional variable importance for random forests, BMC Bioinf. **9**, 307 (2008).

[56] R. Díaz-Uriarte and S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, BMC Bioinf. **7**, 3 (2006).

[57] N. T. Young, G. Allen, J. M. Aiken, R. Henderson, and M. D. Caballero, Identifying features predictive of faculty integrating computation into physics courses, Phys. Rev. Phys. Educ. Res. **15**, 010114 (2019).

[58] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. **7**, 1 (2006), http://jmlr.org/papers/v7/demsar06a.html.

[59] T. G. Dietterich, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Comput. **10**, 1895 (1998).

[60] E. Alpaydm, Combined 5 × 2 cv F test for comparing supervised classification learning algorithms, Neural Comput. **11**, 1885 (1999).

[61] J. M. Aiken, R. De Bin, H. Lewandowski, and M. D. Caballero, Framework for evaluating statistical models in physics education research, Phys. Rev. Phys. Educ. Res. **17**, 020104 (2021).

[62] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018).

[63] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan, Survival ensembles, Biostatistics **7**, 355 (2006).

[64] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, Random forest: A classification and regression tool for compound classification and QSAR modeling, J. Chem. Inf. Comput. Sci. **43**, 1947 (2003).

[65] A. Hapfelmeier, T. Hothorn, K. Ulm, and C. Strobl, A new variable importance measure for random forests with missing data, Stat. Comput. **24**, 21 (2014).

[66] S. v. Buuren and K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, J. Stat. Softw. **45**, 1 (2011).

[67] J. Nissen, R. Donatello, and B. Van Dusen, Missing data and bias in physics education research: A case for using multiple imputation, Phys. Rev. Phys. Educ. Res. **15**, 020106 (2019).

[68] P. Branco, R. P. Ribeiro, and L. Torgo, UBL: An R package for utility-based learning, arXiv:1604.08079.

[69] S. Holm, A simple sequentially rejective multiple test procedure, Scand. J. Stat. Theory Appl. **6**, 65 (1979), https://www.jstor.org/stable/4615733.

[70] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.19.010134 for results related to conditional feature importances, hyperparameter tuning, and Tomek Links. Feature importances for Data Sets 1a and 1b are also included.

[71] L. Hu, J.-Y. J. Lin, and J. Ji, Variable selection with missing data in both covariates and outcomes: Imputation and machine learning, arXiv:2104.02769.

[72] Y. Zhao, Q. Xu, M. Chen, and G. M. Weiss, Predicting student performance in a master of data science program using admissions data, in *Proceedings of The 13th International Conference on Educational Data Mining* (International Educational Data Mining Society, 2020).

[73] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, On the practice of dichotomization of quantitative variables, Psychol. Methods **7**, 19 (2002).

[74] J. R. Irwin and G. H. McClelland, Negative consequences of dichotomizing continuous predictor variables, J. Market. Res. **40**, 366 (2003).

[75] M. Stains *et al.*, Anatomy of STEM teaching in North American universities, Science **359**, 1468 (2018).

[76] K. Commeford, E. Brewe, and A. Traxler, Characterizing active learning environments in physics using network analysis and classroom observations, Phys. Rev. Phys. Educ. Res. **17**, 020136 (2021).

[77] H. Müggenburg, Beyond the limits of memory? The reliability of retrospective data in travel research, Transp. Res. Part A **145**, 302 (2021).

[78] R. Behrens and R. Del Mistro, Analysing changing personal travel behaviour over time: Methodological lessons from the application of retrospective surveys in Cape Town, in *Proceedings of 8th International Conference on Survey Methods in Transport: Harmonisation and Data Quality, Annecy, France* (2008), https://www.researchgate.net/publication/343962276_ANALYSING_CHANGING_PERSONAL_TRAVEL_BEHAVIOUR_OVER_TIME_METHODOLOGICAL_LESSONS_FROM_THE_APPLICATION_OF_RETROSPECTIVE_SURVEYS_IN_CAPE_TOWN_1.

[79] A. L. Edwards, *The Social Desirability Variable in Personality Assessment and Research* (Dryden Press, Ft Worth, TX, 1957), p. viii.

[80] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[81] M. Allen, D. Poggiali, K. Whitaker, T. R. Marshall, and R. A. Kievit, Raincloud plots: a multi-platform tool for robust data visualization [version 1; peer review: 2 approved], Wellcome Open Res. **4**, 63 (2019).

[82] F. J. Massey, Jr., The Kolmogorov-Smirnov test for goodness of fit, J. Am. Stat. Assoc. **46**, 68 (1951).