# Who Gets the Benefit of the Doubt? Racial Bias in Machine Learning Algorithms Applied to Secondary School Math Education

**Haewon Jeong**
Harvard University

**Michael D. Wu**
Harvard University

**Nilanjana Dasgupta**
University of Massachusetts Amherst

**Muriel Médard**
Massachusetts Institute of Technology

**Flavio P. Calmon**
Harvard University

## Abstract

Machine learning algorithms are rapidly being adopted to aid pedagogical decision-making in applications ranging from grading to student placement. Are these algorithms fair? We prove that, for predicting students' math performance, the standard machine learning practice of selecting a model that maximizes predictive accuracy can result in algorithms that give significantly more benefit of the doubt to White, Asian students and are more punitive to Black, Hispanic, Native American students. This disparity is masked by comparatively high predictive accuracy across both groups. We suggest new interventions that help close this performance gap and do not require the use of a different algorithm for each student group. Together, our results suggest new best practices for applying machine learning to math education.

## 1 Introduction

Machine learning (ML) algorithms are routinely used to support decisions that impact millions of students and their educational opportunities. In recent years, ML has been rapidly adopted in areas such as grading (Joyce and Harris, 2018; Newton, 2021), personalized learning (Newton, 2021; Snow, 2019), and school admissions (Newton, 2021; Waters and Miikkulainen, 2014). One of the highest profile applications of algorithmic decision-making to education occurred in 2020 when the UK used a data-driven algorithm to assign 4.6 million grades for the A-level examinations due to Covid-19 restrictions on in-person test taking. This algorithm was found to systematically assign higher grades to students from historically high-performing schools in wealthier regions regardless of students' objective performance and sparked nationwide protests (Quinn, 2020; Coughlan, 2020; Adam, 2020).

ML algorithms are prone to discrimination in domains where racial inequalities are already pervasive. Data-driven algorithms can inherit and exacerbate human biases in applications such as criminal justice (Chohlas-Wood, 2020; Chouldechova, 2017), child welfare (Hurley, 2018; Chouldechova et al., 2018), and hiring (Mann and O'Neil, 2016), resulting in unfair decisions for historically underprivileged groups. In K-12 education—and STEM subjects in particular—racial disparities are widespread, with inequities existing in school funding, access to advanced placement classes, teacher perception, among many other areas (Reardon and Portilla, 2016; Reeves et al., 2016; Gershenson et al., 2016). The persistent racial disparities that exist in K-12 schools create a high-stakes minefield for ML algorithms. Nevertheless, the use of ML in education-related applications continues to increase at an unrestrained pace, with little to no guidelines and best practices to ensure that deployed algorithms are fair to students from diverse backgrounds.

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Workshop on Math AI for Education (MATHAI4ED).

In this work, we demonstrate concrete examples of how racial inequities emerge when ML algorithms are used to predict students' future math performance in secondary education. Our analysis is based on training popular ML models on large-scale datasets collected from middle schools and high schools across the United States. We find that the standard pipeline for training and deploying ML models—collecting a representative dataset, then fitting a ML model to maximize predictive accuracy—can systematically fail Black, Hispanic, and Native-American (BHN) students compared to White and Asian (WA) students when applied to predict future math performance[1].

Accuracy, the metric most often used to choose a ML model, can be deceptive when it comes to assessing whether a ML model is fair in predicting student performance. Accuracy measures the rate of misclassification, but not all errors committed by a ML model are equal. False-positives give the benefit of the doubt to students and provide more opportunities, whereas false-negatives undercut students' future potential. Our experiments show that standard ML algorithms achieve comparably high accuracy for both WA and BHN students when predicting future math performance, but the patterns of misclassifcation can be strikingly different between these two groups: WA students receive substantially more benefit of the doubt while BHN students receive more pessimistic predictions. The harm is silent: the usual procedure of optimizing ML models for accuracy may mask predictions that deprive BHN students of educational opportunities. Our findings suggest that, when predicting student future performance, false-positive rates and false-negative rates across student populations must be closely monitored.

We propose a simple yet effective intervention that significantly reduces the observed gap in false-positive and false-negative rates between BHN and WA students. Naturally, the performance of a ML model depends on the data used to fit the model's parameters, called the *training set*. A standard practice is to use a training set that is representative of the actual student population. We show that by varying mixtures of BHN and WA students in the training set, the gap between error rates can be reduced with minimal impact on the model's overall accuracy. This result indicates that selecting the fairest demographic composition is not always straightforward. In fact, we show that the most counter-intuitive choice of using a homogeneous training set comprised only of one group can result in a model with the smallest disparity between the groups.

## 2 Unequal Benefit of the Doubt

We conducted our analyses on two datasets—the middle school study (MSS) dataset and the public-use high school longitudinal study 2009 (HSLS) dataset (Rogers et al., 2018)[2]. The HSLS dataset is collected from 20,000+ students from 940 high schools across the US. It includes students' and parents' surveys, student demographic information, and students' math performance over the years.

Math is a foundational subject in STEM education. Accurate predictions of students' future math performance can enable better educational resource distribution to boost students with potential (e.g., advanced class placements or gifted program recommendations). We train a ML model to predict if a student will be a top 50% performer in their future math class (positive prediction) or bottom 50% performer (negative prediction). Past performance is not sufficient to predict future success and persistence. With the HSLS dataset, we observe that prediction accuracy is $68.2 \pm 0.1$ % if we predict students' math performance in the 9th grade based only on their past performance. Accuracy improves to $75.0 \pm 0.1$ %, by utilizing more features such as students' and parents' survey responses.

Being able to take advantage of more data for more accurate predictions with ML sounds promising. However, the deployment of ML may not benefit all racial groups equally. The models we trained achieve comparable accuracy across WA and BHN students. However, when we examine metrics beyond accuracy, significant racial inequalities emerge. Even though a similar accuracy indicates that there are roughly equal numbers of misclassified points among WA and BHN students, *how they are misclassified is staggeringly different.*

We examine four different metrics: accuracy, false positive rate (FPR), false negative rate (FNR), and predicted base rate (PBR). False positive prediction refers to students who did not belong in the top 50% of math performers based on their actual grades but received a positive prediction from the ML

[1] We perform this binary categorization of races since BHN students are historically underrepresented in STEM education (National Science Foundation, 2017).

[2] Full results including the MSS dataset are included in the appendix.

| | | WA | BHN | Difference |
|---|---|---|---|---|
| HSLS | Size | 3318 | 1695 | – |
| | PBR | **0.580** ±0.002 | 0.347±0.002 | 0.233 (+40.17%) |
| | FPR | **0.304** ±0.003 | 0.176±0.002 | 0.128 (+42.11%) |
| | FNR | 0.209±0.002 | **0.371** ±0.004 | -0.162 (-77.51%) |
| | Accuracy | 0.750±0.001 | 0.750±0.002 | 0.000 (+0.00%) |

Table 1: Random forest results for math performance predictions.

model. In other words, they are given the *benefit of the doubt* from the ML prediction. On the flip side, false negative predictions are students who did belong in the top 50% in reality, but are given a negative prediction. This is a *pessimistic underestimation* of their future performance.

**High school dataset results.** We trained several widely-used binary classification models (logistic regression, SVM, random forests)[3] to predict top and bottom 50% performers in the standardized test taken in the 9th grade. We removed all features related to students' race such as their parents' place of birth. We trained different models with 30 different train/test splits and obtain the average and standard error. Since random forest showed the best accuracy among the three models, we only present random forest results in Table 1 (full results are in Appendix C).

First, notice that the difference in accuracy between WA and BHN is negligible. However, FNR was considerably smaller for WA students compared to BHN. The relative difference in FNR was up to 78%. This implies that WA students are less prone to get an underestimated prediction by the ML model. At the same time, FPR is 42% higher for WA than BHN students. In other words, WA students more frequently receive the benefit of the doubt from the trained ML model. We also observe that PBR is higher in WA students than in BHN students. This may reflect the difference in the ground truth data. The observed base rate was 0.57 for WA and 0.38 for BHN students. (difference = 0.19). However, the PBR difference from the trained random forest models was about 0.23, indicating that the existing racial performance gap is exaggerated in the ML model's predictions.

By examining FPR and FNR, we discover that WA students are consistently given more benefit of the doubt, while BHN students are consistently underestimated in predicting their future math performance despite similar accuracy numbers for both groups. This shows that narrowly focusing on accuracy can give an illusion of fairness when there is significant discriminatory impact on students from historically underrepresented groups. Removing implicitly race-related features through a careful analysis reduces the gaps in FPR and FNR, but significant gaps still remain (see Appendix B).

**Implications of FNR and FPR gaps.** When predicting student performance, unequal error rates have real-world consequences. Consider a scenario where we use a random forest model trained on the HSLS dataset for 9th grade math placement. Students who are predicted to be in top 50% will be placed in the advanced-level math class and students who receive bottom 50% prediction will be placed in the basic math class. The FPR of 0.30 for WA students (see Table 1) means that 30% of the students who would not perform well in the 9th grade will be placed in the advanced class. They are given the benefit of the doubt and the opportunity to learn more advanced math. On the other hand, only 18% of the BHN students get the same benefit of the doubt (FPR=0.18). The FNR of 0.21 in WA students indicates that 21% of WA students who would in fact perform well in the future will be placed in the basic class by the ML algorithm. For BHN students, a startlingly high 37% will be incorrectly placed in the basic class, their academic potential ignored by the algorithm.

The downstream effects of such misclassification is disproportionately detrimental to BHN students. Missing the opportunity to take foundational math classes such as Algebra 1 can prevent them from taking further advanced classes in the following years. Indeed, past research shows that middle school algebra is a strong early predictor of educational outcomes in high school and college (Berwick, 2019; Loveless, 2013; Maltese and Tai, 2011; Stein et al., 2011). Moreover, they are at risk of losing interest in STEM subjects because of the pessimistic prediction by the algorithm. It was shown in prior research that low test scores or class placements to less-advanced classes discourage students from historically marginalized groups more because they elevate negative stereotypes (Riegle-Crumb et al.,

[3]See Appendix A for details.

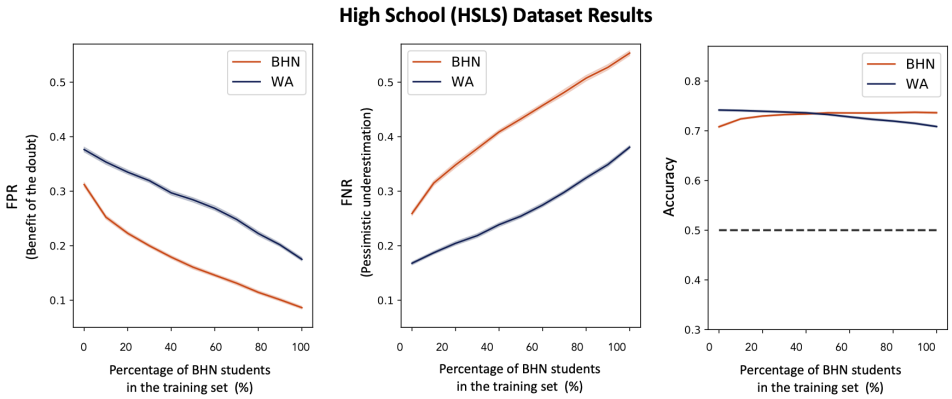


Figure 1: Results of changing racial composition of training sets.

2019; Sanabria and Penner, 2017). In the fair ML literature, the importance of equal FPR and FNR was recognized, and well-known fairness metrics such as equalized odds and equalized opportunity are based on this notion. In other words, what we show in this paper is that in terms of equalized odds, vanilla ML algorithms perform very poorly across the racial groups (see Appendix B).

## 3 Racial Composition of Training Set and Fairness

For the results we presented in the previous section, we did not balance racial groups as the given dataset is already a nationally representative sample, which is collected from both public and private schools sampled according to national demographics. Was it *fair* to use the HSLS dataset as it is? Or, is it more *fair* to rebalance the dataset to have the equal number of data points in each group? When dealing with a dataset made up of different population groups, whether to balance the dataset and how to balance the dataset are unavoidable choices that a data scientist has to make in the data preprocessing stage of the ML pipeline. Despite its importance, past research has not rigorously investigated what racial composition of training sets would produce the most fair model.

To investigate this question, we trained ML models with different racial mixtures in the training set. In the experiments, we varied the proportion of BHN students in a training set, $p = \frac{(\#\text{BHN})}{(\#\text{WA})+(\#\text{BHN})}$, from 0 to 1, in the interval of 0.1. For all $p$, we ensure that the entire training set size (i.e., number of all data points in the training set) is the same. To achieve different $p$, we subsample from each group and fit a random forest classifier for each subsampled dataset. The results are summarized in Fig. 1.

The results we observe are striking. Focusing solely on accuracy may lead to the incorrect conclusion that the effect of different racial compositions of a training set is minute: the accuracy for each group does not vary more than 0.05 as we change $p$ from 0 to 1 (i.e., 0% to 100% BHN). However, FPR and FNR metrics change drastically with different racial compositions of the training set. FPR monotonically decreases and FNR monotonically increases for both BHN and WA students as we increase $p$ from 0% BHN to 100% BHN. The range of FPR difference is from $\sim$0.4 to 0.1 and FNR moves from 0.2 to 0.5. The gaps in FPR and FNR remain throughout different values of $p$, but they reduce slightly around $p = 0$.

## 4 Conclusion and Discussion

Together, our findings are of direct value to data scientists, school administrators, and teachers who are considering using ML to support pedagogical decisions. The results presented next suggest two critical best practices: 1) monitor differences between false-positive and false-negative rates across student groups to ensure that all students receive comparable benefit of the doubt regardless of their racial background, and 2) judiciously vary the racial composition of training sets in order to close the gap between false-positive and false-negative rates. One important future direction that has to be examined is how existing fair intervention methods would perform on education data.

## References


Adam, K. (2020). The U.K. used an algorithm to estimate exam results. the calculations favored elites. https://www.washingtonpost.com/world/europe/the-uk-used-an-algorithm-to-estimate-exam-results-the-calculations-favored-elites/2020/08/17/2b116d48-e091-11ea-82d8-5e55d47e90ca_story.html.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR.

Berwick, C. (2019). Is it time to detrack math? https://www.edutopia.org/article/it-time-detrack-math.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328.

Chohlas-Wood, A. (2020). Big data's disparate impact. https://www.brookings.edu/research/understanding-risk-assessment-instruments-in-criminal-justice.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148. PMLR.

Coughlan, S. (2020). A-level results: 'huge mess' as exams appeal guidance withdrawn. https://www.bbc.com/news/education-53795831.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Gershenson, S., Holt, S. B., and Papageorge, N. W. (2016). Who believes in me? the effect of student–teacher demographic match on teacher expectations. *Economics of education review*, 52:209–224.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Hurley, D. (2018). Big data's disparate impact. https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html.

Johndrow, J. E. and Lum, K. (2019). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220.

Joyce, J. L. and Harris, L. (2018). Artificial intelligence (AI) and education. *Focus, Congressional Research service, August.*

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., and Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*, pages 853–862.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015). The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Loveless, T. (2013). The 2013 brown center report on american education: How well are american students learning? Technical report, The Brown Center on Education Policy at the Brookings Institution.

Maltese, A. V. and Tai, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in stem among us students. *Science education*, 95(5):877–907.

Mann, G. and O'Neil, C. (2016). Hiring algorithms are not neutral. *Harvard Business Review*, 9:2016.

National Science Foundation (2017). National center for science and engineering statistics, women, minorities, and persons with disabilities in science and engineering.

Newton, D. (2021). From admissions to teaching to grading, AI is infiltrating higher education. https://hechingerreport.org/from-admissions-to-teaching-to-grading-ai-is-infiltrating-higher-education/.

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

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. *arXiv preprint arXiv:1709.02012*.

Quinn, B. (2020). Uk exams debacle: how did this year's results end up in chaos? https://www.theguardian.com/education/2020/aug/17/uk-exams-debacle-how-did-results-end-up-chaos.

Reardon, S. F. and Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *Aera Open*, 2(3):2332858416657343.

Reeves, R., Rodrigue, E., and Kneebone, E. (2016). Five evils: Multidimensional poverty and race in america. *Economic Studies at Brookings Report*, 1:1–22.

Riegle-Crumb, C., King, B., and Irizarry, Y. (2019). Does stem stand out? examining racial/ethnic gaps in persistence across postsecondary fields. *Educational Researcher*, 48(3):133–144.

Rogers, J. E., Ritchie, E., and Fritch, L. B. (2018). Hsls:09 base year to second follow-up public-use data file. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2018142.

Sanabria, T. and Penner, A. (2017). Weeded out? gendered responses to failing calculus. *Social Sciences*, 6(2):47.

Snow, J. (2019). AI technology is disrupting the traditional classroom. here's a progress report. https://www.pbs.org/wgbh/nova/article/ai-technology-is-disrupting-the-traditional-classroom/.

Stein, M. K., Kaufman, J. H., Sherman, M., and Hillen, A. F. (2011). Algebra: A challenge at the crossroads of policy and practice. *Review of Educational Research*, 81(4):453–492.

Waters, A. and Miikkulainen, R. (2014). Grade: Machine learning support for graduate admissions. *AI Magazine*, 35(1):64–64.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.

# A Experiment Details

**Datasets** In all analyses we present in the paper, we use two datasets: the middle school study (MSS) dataset[4] and the high school longitudinal study 2009 (HSLS) dataset.

The HSLS dataset is collected from 20,000+ students from 940 public and private high schools in the US across 50 different states and the District of Columbia. The first data collection was in 2009 when the students were in their 9th grade fall term, first follow-up was collected in 2013 at the end of high school, and the second follow-up was in 2016, approximately 3 years after the high school completion. The dataset contains more than 3,000 features. We picked a set of 52 variables that closely resemble the variables in the MSS dataset. They are composed of 31 parent survey features, 20 student survey features, and students' math grade in the most advanced class they took in the 8th grade. More details below:

- Student variables: (i) student demographics (race, ethnicity, gender); (ii) self-efficacy in math; (iii) student aspiration in future math classes; (iv) student beliefs and stereotypes on math (e.g., "Do you agree that if you spend a lot of time and effort in your math and science classes, you won't be popular?");
- Parent and family characteristics: (i) parents' social class; (ii) parents' profession (STEM or non-STEM); (iii) aspirations for their child; (iv) perceptions of child's talent in STEM; (v) STEM related enrichment activities they engage in with their child.
- Performance variables: (i) 8th grade end of year final math grade; (ii) standardized math test score in 9th grade; (iii) standardized math test score in 12th grade.

**Method** We train ML algorithms for binary classification on students' 9th grade math performance. We train three classes of ML models that are commonly used for classification: random forest, support vector machine (SVM), and logistic regression. As random forest showed the best accuracy in all prediction tasks, we only present random forest results for brevity. The full results on all three models are given in the SI. We use Scikit-learn (Pedregosa et al., 2011) package to implement the three models. We use the same hyperparameters for all the experiments in the paper as they showed good accuracy in all of the prediction tasks after hyperparameter tuning. They are summarized below:

- Random forest: max depth=16, min samples per leaf=3, number of trees = 100
- SVM: RBF kernel, C = 1.0
- Logistic regression: L2 regularizer, C = 1.0

We evaluate our models in four metrics, accuracy, false positive rate (FPR), false negative rate (FNR), and predicted base rate (PBR), which are defined as follows:

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{All}},$$

$$\text{FPR} = \frac{\text{False Positive}}{\text{True Negative} + \text{False Positive}},$$

$$\text{FNR} = \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}},$$

$$\text{PBR} = \frac{\text{True Positive} + \text{False Positive}}{\text{All}}.$$

For the experiments presented in Table 1, the training set size of the HSLS dataset was 10,156. For the racial composition experiments shown in Fig. 1, the training set size was 4,120 for the HSLS dataset. Note that training set size reduces substantially to accommodate the case for $p = 1.0$, i.e., we only use BHN students for training. For each $p \in \{0.0, 0.1, 0.2, \cdots, 1.0\}$, we trained a random forest classifier with 30 (10) different subsampling runs and 30 (10) different train-test splits for the MSS (HSLS) dataset.

For each metric, we compute standard error by computing the standard deviation of the metrics we obtain from all the models we trained and then dividing it by $\sqrt{n}$, where $n$ is the number of trained models.

[4] Deatails about the MSS dataset will be released upon acceptance.

| | WA | BHN | Difference |
|---|---|---|---|
| Size | 3278 | 1736 | – |
| PBR | **0.558** ±0.002 | 0.378±0.002 | 0.180 (+32.26%) |
| FPR | **0.279** ±0.003 | 0.205±0.002 | 0.074 (**+26.52%**) |
| FNR | 0.229±0.002 | **0.339** ±0.003 | -0.110 (**-48.03%**) |
| Accuracy | 0.749±0.001 | 0.744±0.002 | 0.005 (+0.67%) |

Table 2: Random forest results without using race predicting features on the high school (HSLS) dataset. This is the result of running the same task of predicting 9th grade math performance, but after removing five features that are most related to students' race. Even though the gaps in PBR, FPR, and FNR remain, we notice that FPR gap of 42% in Table 1 reduces to 27%, and similarly FNR gap reduces from 78% to 48%.

## B Omitted Analyses and Discussions

**Connection to fairness metrics in the ML literature**

The four metrics (PBR, FPR, FNR, and accuracy) we evaluate throughout the paper are related to some of the widely-used fairness metrics in ML: statistical parity (Dwork et al., 2012) and equalized odds/opportunity (Hardt et al., 2016). Consider two population groups, a minority group (group 0) and a majority group (group 1). When an ML model has the same PBR for group 0 and group 1, it satisfies *statistical parity*. When the model has the same FNR and FPR (i.e., $\text{FPR}_0 = \text{FPR}_1$ and $\text{FNR}_0 = \text{FNR}_1$), it satisfies the *equalized odds* criterion. A relaxed version is *equalized opportunity*, which only requires equal FNRs: $\text{FNR}_0 = \text{FNR}_1$. These metrics have been heavily analyzed on datasets from domains such as criminal justice, income prediction, and healthcare (Hardt et al., 2016; Krasanakis et al., 2018; Pleiss et al., 2017; Hardt et al., 2016; Celis et al., 2019; Donini et al., 2018; Agarwal et al., 2018), but to the best of our knowledge, such evaluation has not been reported on K-12 education data. Our results clearly illustrate that significant differences in equalized odds metrics can also arise when off-the-shelf ML algorithms are applied on secondary school student data.

**Mechanism for unequal benefit of the doubt**

How does a ML model systematically underestimate BHN students' performance and overestimate WA students' performance despite not using any race-related features to make predictions? One possible hypothesis is that there are other features in the data that are covariates for students' race. The trained ML model can then exploit these features to assign disparate predictions to different racial groups.

To test this hypothesis, we designed the following experiment. First, we identify if any features reveal information about students' race by training an ML model that performs a new binary classification task of predicting whether a student is WA or BHN on the HSLS dataset. If it performs better than random guessing, we can conclude that other features in the data predict students' race. In our case, baseline accuracy of random guessing is 0.66 as a model that predicts everyone is WA achieves the accuracy of 0.66 as 66% of the population is WA. A random forest model we trained achieved the accuracy of 71% accuracy. We then ranked the most relevant features used in the prediction to infer the most race-revealing features. The five most predictive features were: `S1LANG1ST` (student's first language), `P1MARSTAT` (parent 1's marital status), `X1FAMINCOME` (family income), `X1PAR2EDU` (parent 2's highest level of education), and `X1PAR2OCC2` (parent 2's current/most recent occupation). Then, we trained a new model without using these five features, i.e., with 47 features instead of 52. If the racial gap reduces by removing the most race-related features, it supports our hypothesis that implicitly race-related features were being used to assign different predictions to different races. We present the result of training a random forest model with the reduced feature set in Table 2. We observe that gaps in PBR, FPR, and FNR all decrease substantially while maintaining a similar accuracy.

This shows that removing race information from the training data is not enough to prevent racially discriminatory performance of an ML model, but a careful feature selection can considerably reduce

racial gaps. This observation is consistent with reports of ML bias in other applications such as criminal recidivism prediction (Johndrow and Lum, 2019). One can also employ preprocessing techniques that reduce race-related information in the data while maintaining the useful information for prediction (Zemel et al., 2013; Feldman et al., 2015; Louizos et al., 2015).

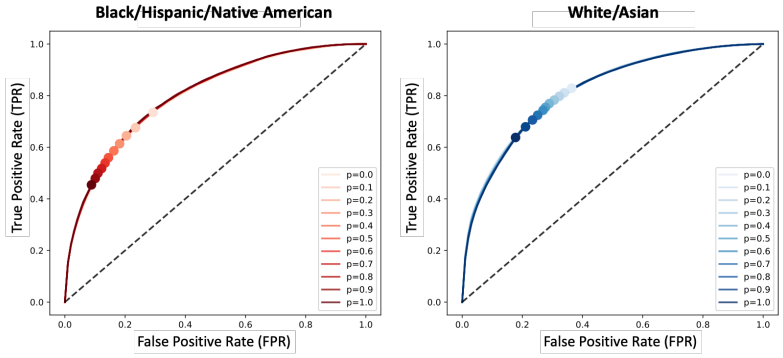


Figure 2: **ROC curve analysis of training with different racial mixtures on the HSLS dataset.** We plot receiver operating characteristic (ROC) curves for each different mixture (i.e., each different $p$). Lighter colors represent smaller $p$ (i.e., fewer BHN students in the training set) and darker colors represent bigger $p$. Eleven different ROC curves almost overlap and are indistinguishable from each other. The circle markers denote the operating point of the trained model with different values of $p$. Notice that as $p$ increases, the markers move down on the curve and have smaller FPR and TPR values. This implies that training with different racial mixtures mimics the behavior of choosing different decision thresholds of a fixed classifier.

### ROC Curve Analysis

There is a clear trade-off between FPR and FNR as we change the racial mixture of the training set. This behavior can be mapped to the receiver operating characteristic (ROC) curve of the classifiers used to predict student math performance. For binary classification, the ML models we considered produce a *score* $S$ ($0 \leq S \leq 1$) for each input sample. Predictions are then made by thresholding the score as follows:

$$\text{Prediction} = \begin{cases} \text{Positive}, & \text{if } S \geq 0.5, \\ \text{Negative}, & \text{if } S < 0.5. \end{cases}$$

After training, a ML model can be viewed as a function that computes a score $S$ from the input data. In most off-the shelf models, the default score threshold for predicting positive outcomes is 0.5, but this threshold can be adjusted. Increasing the threshold above 0.5 leads to fewer data points receiving positive predictions and, equivalently, a lower FPR and a higher FNR. When predicting student performance, this corresponds to fewer students being flagged as high performers and less benefit of the doubt overall. Conversely, lowering the threshold below 0.5 results in a lower FNR and a higher FPR. ROC curves show how true positive rate (TPR) and FPR change as we vary the threshold for a trained classifier (i.e., for a fixed function that computes $S$ from the input), where TPR is defined as:

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = 1 - \text{FNR}.$$

Notice that changing a score threshold for a trained classifier trades off FPR and FNR, similar to what we observed when changing the racial mixture of the training set. Is changing the racial composition of the training set equivalent to changing the score threshold? The goal of following analysis is to understand if models trained on different racial mixtures produce different FPR and FNR because they have different ROC curves or if they approximately learn the same scoring function (i.e., have the same ROC curves) but use different thresholds (i.e., correspond to different points on the ROC curve).

We plot the ROC curves for the trained models for each $p$ in Fig. 2. Recall that $p$ is the proportion of BHN data points in the training set. The lighter color lines and markers in the plot represent smaller

$p$. We observe that the eleven ROC curves for different values of $p$ mostly overlap with insignificant differences. The markers in the plot represent the operating point for the trained classifier for each $p$. As we increase $p$, these points move down on the curve. This is essentially equivalent to increasing the decision threshold of a fixed classifier to be more conservative in making positive predictions. In other words, the classifier decreases FPR and increases FNR.

We now explain this behavior with the base rate change induced by different racial mixture. Base rate (BR) is defined as:

$$\text{BR} = \frac{\text{True Positive} + \text{False Negative}}{\text{All}}.$$

As we increase $p$ to include more BHN students, the overall BR of the training set becomes smaller. The classifier that maximizes accuracy on the ROC curve corresponds to the point that has the tangent of $\frac{1-\text{BR}}{\text{BR}}$. Since BR decreases with increasing $p$, the $\frac{1-\text{BR}}{\text{BR}}$ becomes larger. Hence, the operating point moves down on the ROC curve where the slope is steeper.

**Best practices**

Our analyses on the middle school and high school datasets show that even when an algorithm gives significantly more benefit of the doubt to the privileged groups of students compared to historically marginalized groups, accuracy for the two groups can be almost equal. This result serves as a cautionary tale on the danger of blindly following the standard practice of choosing a model that achieves high accuracy. To build a fair algorithm, it is necessary to also examine FPR and FNR metrics to identify which students are receiving the benefit of the doubt and whose academic potential is being underestimated by the algorithm. Algorithm developers working in education must give FPR and FNR metrics as much importance as accuracy and report FPR and FNR metrics for each racial group.

When a data scientist observes unequal benefit of the doubt between racial groups in ML models, they can consider using intervention methods that reduce the FPR and FNR gaps. As mentioned earlier, equal FPR and FNR is referred to as *equalized odds* in the fair ML literature (Hardt et al., 2016). This notion of group fairness has been studied extensively and various methods have been proposed to reduce FPR/FNR disparities, ranging from preprocessing techniques (Krasanakis et al., 2018), postprocessing techniques (Pleiss et al., 2017; Hardt et al., 2016), to adding FPR/FNR equality as an objective during the training process (Celis et al., 2019; Donini et al., 2018; Agarwal et al., 2018).

Our experiments of altering the demographic composition in training sets add a new dimension to fairness-ensuring interventions in ML. By using a training set with different ratios of population groups, we arrive at different models which can improve FPR/FNR disparities with little to no sacrifice in accuracy. By doing an analysis similar to Fig. 1, a data scientist can select a training set that improves the benefit of the doubt given to all groups. The main advantages of this method is that it does not require deploying different models for different groups (or using race as an input feature) nor any change to the data beyond varying the composition of the training set. This intervention can also easily be paired with other existing fair learning algorithms later in the ML pipeline.

## C Omitted Results

Table 3 presents the results from SVM and Logistic Regression trained models for the HSLS dataset, as well as the Random Forest model results presented in the paper. Similar to the middle school dataset, note that the difference in accuracy between all three models is minute (less than 2% between WA and BHN groups) and how the top-line accuracy numbers are quite high for all three models (74%-75%). And yet despite these high accuracy numbers, peeling back the model's performance by looking at its mistakes shows significant disparities in FPR and FNR rates between historically advantaged WA groups and historically disadvantaged BHN groups. Consider the False Negative Rate (FNR), or the model's underestimations of student performance. For Random Forest models, the Black, Hispanic, and Native American students have a shocking 77.51% higher rate of underestimation, and it's similar for SVM and Logistic Regression models (78.64% and 61.40% respectively). Looking at False Positive Rate (FPR) paints a similar story, where WA students are consistently given more benefit of the doubt than BHN students. Similar to the MSS dataset, we observe a fairly large gap in PBR between WA students and BHN students. Again, this can reflect differences in the ground truth

Table 3: HSLS Race Results (unbalanced, no sensitive attributes, with grade8th)

| | | WA | BHN | Difference |
|---|---|---|---|---|
| Random Forest | Size | 3318 | 1695 | – |
| | PBR | 0.580±0.002 | 0.347±0.002 | 0.233 (+40.17%) |
| | FPR | 0.304±0.003 | 0.176±0.002 | 0.128 (+42.11%) |
| | FNR | 0.209±0.002 | 0.371±0.004 | -0.162 (-77.51%) |
| | Accuracy | 0.750±0.001 | 0.750±0.002 | 0.000 (+0.00%) |
| SVM | Size | 3318 | 1695 | – |
| | PBR | 0.585±0.002 | 0.336±0.002 | 0.249 (+42.56%) |
| | FPR | 0.330±0.002 | 0.171±0.002 | 0.159 (+48.18%) |
| | FNR | 0.220±0.002 | 0.393±0.004 | -0.173 (-78.64%) |
| | Accuracy | 0.732±0.001 | 0.745±0.002 | -0.013 (-1.78%) |
| Logistic | Size | 3318 | 1695 | – |
| | PBR | 0.576±0.002 | 0.360±0.002 | 0.216 (+37.50%) |
| | FPR | 0.321±0.003 | 0.195±0.002 | 0.126 (+39.25%) |
| | FNR | 0.228±0.002 | 0.368±0.004 | -0.140 (-61.40%) |
| | Accuracy | 0.732±0.001 | 0.739±0.002 | -0.007 (-0.96%) |

Table 4: MSS Race Results (balanced, no sensitive attribute, with grade8th)

| | | WA | BHN | Difference |
|---|---|---|---|---|
| Random Forest | Size | 128 | 132 | – |
| | PBR | 0.575±0.008 | 0.488±0.008 | 0.087 (+15.13%) |
| | FPR | 0.319±0.011 | 0.281±0.010 | 0.038 (+11.91%) |
| | FNR | 0.205±0.011 | 0.276±0.009 | -0.071 (-34.63%) |
| | Accuracy | 0.740±0.006 | 0.721±0.007 | 0.019 (+2.57%) |
| SVM | Size | 128 | 132 | – |
| | PBR | 0.571±0.008 | 0.511±0.008 | 0.060 (+10.51%) |
| | FPR | 0.408±0.011 | 0.398±0.012 | 0.010 (+2.45%) |
| | FNR | 0.289±0.011 | 0.363±0.011 | -0.074 (-25.61%) |
| | Accuracy | 0.654±0.006 | 0.618±0.008 | 0.036 (+5.50%) |
| Logistic | Size | 128 | 132 | – |
| | PBR | 0.580±0.007 | 0.463±0.007 | 0.117 (+20.17%) |
| | FPR | 0.398±0.011 | 0.305±0.009 | 0.093 (+23.37%) |
| | FNR | 0.264±0.010 | 0.356±0.012 | -0.092 (-34.85%) |
| | Accuracy | 0.673±0.007 | 0.670±0.006 | 0.003 (+0.45%) |

data – the actual ground truth base rate was 0.57 for WA and 0.38 for BHN (a difference of 0.19). But consistently in all three models, the PBR is 0.23 for Random Forest, 0.25 for SVM, 0.22 for Logistic Regression, the existing racial performance gap is exaggerated in the ML model.

Table 4 presents the results from SVM and Logistic Regression trained models for the MSS dataset, as well as the Random Forest model results presented in the paper. First, notice that the difference in accuracy between WA and BHN is minute across all three models. The relative difference is less than 5%. However, in all three models, FNR was considerably smaller for WA students compared to BHN. The relative difference was up to 35% in random forest and logistic regression and 26%

in SVM. This implies that WA students are less prone to get an underestimated prediction by the ML model. At the same time, FPR is higher for WA students, 12% higher in random forest and 23% higher in logistic regression (no significant difference in SVM). In other words, WA students are more frequently offered the benefit of the doubt from the trained ML model. We also observe that PBR is consistently higher in WA students than in BHN students. This can reflect the difference in the ground truth data. The actual ground truth base rate was 0.52 for WA and 0.46 for BHN students, making the difference of 0.06. However, in random forest models the PBR difference was about 0.09 and it was 0.12 in logistic regression model, which means that the existing racial performance gap is exaggerated in the ML model.