Fairness in AI: A Numerical Analysis on Student Performance in Secondary School

Danielle Sim
Applied Data Science, MS
University of Southern California
simd@usc.edu

Saurabh Jain
Applied Data Science, MS
University of Southern California
sjain681@usc.edu

Abstract

The primary goal of the project is to investigate possible bias in a dataset and examine how such bias affects a prediction model. The dataset at hand comes from the University of California, Irvine's Machine Learning Repository, called the 'Student Performance Data Set' [2] The prediction model goal is to predict students' final math grades which is on a 0 to 20 scale. We aim to understand the bias in data by investigating selected protected features such as sex, and the bias in prediction based on training multiple linear regression models to predict final math grades. Finally, we will explore some ways to mitigate bias in our predictions and explore how different definitions and implementations of fairness affect the results.

1 Introduction

There are many public school programs across the United States that are designed in some way for "gifted" or "talented" students (ex: GATE Program in California) to be identified in early years such as elementary school, and given more funding, attention, and education. How these students are identified has changed over the years, whether it be through teacher recommendations, IQ tests, or other measures, standardized or not. It is not unfeasible to imagine a public school or district to implement some form of Machine Learning or AI to automatically identify such students, thinking that it is a "more fair" method of selecting these students rather than basing it off subjective opinions of their teachers. However, datasets, despite containing "objective" facts, still contain bias, and the models that are built off these datasets with goal of either predicting student behavior or identifying certain groups of students have high potential to learn and further perpetuate biases.

The primary goal of this project is to investigate such possible bias in a dataset and examine its effects on prediction modeling. Other objectives include implementing methods that may mitigate bias and determine if a fairer model can be achieved. The dataset 'Student Performance Data Set' comes from secondary education data of two schools in Portugal and contains 382 observations or students and 33 attributes including sex, age, family size, address (urban or rural), parent demographic information, and other information such as student health, absences, and extra curricular activities. The diverse set of features that include demographic, social, and school related information for each student will allow for an extensive analysis on potential confounding variables and correlations with the prediction outcomes on final Portuguese and Math grades.

Several ways to mitigate bias will be explored, such as data augmentation, fairness through unawareness, group fairness or statistical parity in which the distribution of good outcomes are the same between two groups of a protected feature, conditional statistical parity, equalized odds, and equality of opportunity. We will also explore methods to achieve fairness on the individual level such as implementing counterfactual fairness, and lastly fairness on the subgroup level. For each

Preprint. Under review.

implementation of these definitions of fairness, prediction models to predict final class grades will be constructed and to examine and compare accuracy and fairness.

2 Related Work

Studies have been done to determine features that can predict student academic performance irrespective of their level of study. One such study is "Using Data Mining to predict Secondary School Performance" [2]. It confirms the conclusion found in Predicting Students' Performance in Distance Learning Using Machine Learning Techniques [5]: student achievement is highly affected by previous performances. Nevertheless, an analysis comparing the best predictive models has shown that there are other relevant features such as: school related (e.g. number of absences, reason to choose school, extra educational school support), demographic (e.g. student's age, parent's job and education) and social (e.g. going out with friends, alcohol consumption) that best predict academic performance.

Meanwhile, a case study of Machakos teachers college portrays that students admitted with high grades are expected to outperform those admitted with lower grades but this was not the result; groups of students who had received C+ in KCSE (Kenya Certificate of Secondary Education) performed approximately the same as those who had scored a C across subjects [1]. This indicates that the student admission grade into the college does not count in their performance in the Primary teacher education curriculum. Similarly the students' age and gender are not contributory factors in their academic achievement according to this study. Rather, focus and preparedness determine good performance regardless of these demographic attributes.

Various studies have found that students from the least affluent socio-economic groups tend to perform less well than their more affluent peers [10]. Finally, personal characteristics such as sex and ethnicity are also known to influence academic performance. Though the present study does not focus on ethnicity, significant differences in performance and participation have been documented between ethnic groups. Another study confirms that by treating student performance in the core knowledge courses as an early warning signal, faculty, administrators, and students might be able to identify students who could benefit from early intervention and help them increase their probabilities of academic success in business studies [4]. While these studies' primary focuses have been on predicting academic performance with respect to specific features such as demographic attributes or previous academic history, in this study we will be searching for any source of bias in predicting academic performance with respect to protected features such as student age.

3 Preliminary Analysis

3.1 Data Preprocessing and Analysis Plan

The dataset provided by UCI's Machine Learning Repository comes as two separate datasets: one for student grades in Portuguese, a language class, and another for grades in the same group of students' math class grades. Both datasets contain student demographic and other socioeconomic information and the 2 datasets were merged. Features of sex and age were pre-identified as protected features of interest, and mapped to binary features where students younger than 18 have an age binary value of 0, and students 18 and older have an age binary value of 1. Males and Females were mapped to 1s and 0s, respectively.

The analysis plan consists of predicting Portuguese and Math grades for the students, while looking at the 2 different protected features. Predictive analysis would be performed twice, once to predict Portuguese grades, and one to predict Math grades, and for each prediction, fairness metrics would be performed twice, once with respect to sex, and again with respect to age. To do this, grades are binned into binary variables, where final grades less than 10 were given a value of 0, and final grades 10 and above given value of 1.

3.2 Preliminary Analysis and Exploration

Preliminary exploratory analysis was done both to explore the data and also as an initial look into the fairness of the dataset. The dataset was found to be balanced with respect to gender, with 198 female students and 184 male students. Final Math and Portuguese grade distributions can be seen in Table 1

as the outcome variable and tabulating with respect to sex, the selected protected feature. A similar analysis was conducted to explore Age (older vs younger students) as a second potential protected feature, but later not considered in analysis, as sex turned out to be a more interesting and relevant feature to select in this analysis. Histograms of these distributions can be seen in Figure 1.

Class	Sex	Min	Q1	Median	Mean	Q3	Max
Portuguese	Females	0	12	13	13.137	15	19
	Males	0	10	12	12.039	14	19
	All Students	0	8	12	12.516	15	19
Math	Females	0	8	10	10.059	13	19
	Males	0	9	11	11.059	14	20
	All Students	0	8	11	11.902	14	20

Table 1: Final Grade Distributions

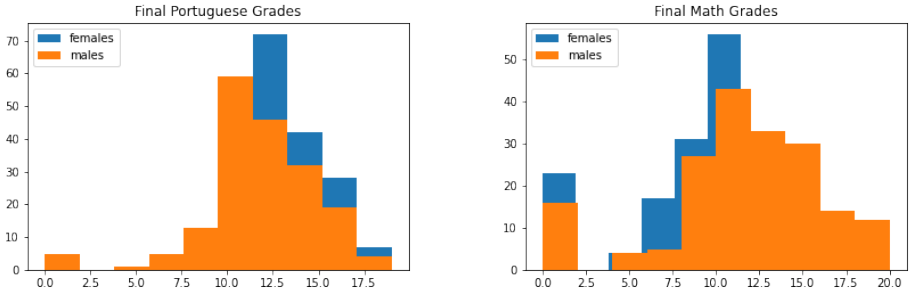

Figure 1: Final Grade Distributions

From this exploratory analysis it is notable that female students had a higher final grade average in Portuguese than their male student counterparts, while male students had a higher final grade average in Math than their female student counterparts.

While final grades are a continuous feature, the feature was binned in a binary variable such that a logistic regression could later be performed. Feature counts were evaluated and results can be seen in Tables 2 and 3.

	Female	Male	All Students
Low	26	52	79
High	172	132	304
All Students	198	184	382

Table 2: Cross-Tab of Final Portuguese Grades (binned) by Sex

	Female	Male	All Students
Low	105	78	183
High	93	106	199
All Students	198	184	382

Table 3: Cross-Tab of Final Math Grades (binned) by Sex

T-tests were performed to evaluate for statistical significance in the differences between groups in this dataset. This was done with respect to both mean of the numeric grade and also mean of the binary variable grade. The t-test p-values can be seen in Table 4. Looking at Portuguese grades with respect to sex had statistically significant p-values in the t-tests in both numeric and binary formats. This was a strong motivator in selecting sex as the protected feature to investigate in this analysis. Math in this comparison is not significant. When looking at grades with respect to age as a protected feature, only Math grades are significant as a binary feature.

Format	Feature	Sex	Age
Numeric	Portuguese	0.0007	0.224
	Math	0.063	0.084
Binary	Portuguese	0.0003	0.804
	Math	0.074	0.047

Table 4: T-Test p-values for class grades with respect to protected features

4 Analysis Methods and Results

4.1 Full Logistic Regression

A full logistic regression analysis was done to establish a baseline performance before implementing any methods for feature selection or to mitigate bias for predicting Portuguese and Math grades (binary). Using all features to predict Portuguese grade, the full logistic regression model has accuracy of 79.22%. Statistical parity and equal opportunity were also calculated with these predictions with respect to sex. Repeating the analysis for Math, using all features to predict Math grade, the full logistic regression model has accuracy of 59.74%, and again statistical parity and equal opportunity were calculated with these model predictions with respect to sex. Formulas for statistical parity and equal opportunity are based on metrics described in Mehrabi et. al. [7]. These results can all be seen in Table 5.

Model	Accuracy	Statistical Parity	Equal Opportunity
Portuguese (Full)	79.22%	-0.152	-0.118
Math (Full)	59.74%	0.356	0.382
Portuguese (RFE)	77.92%	-0.172	-0.118
Math (RFE)	63.63%	0.203	0.190

Table 5: Model Accuracy and Fairness Metrics

4.2 Feature Selection

To identify important variables in the dataset to perform logistic regression, Recursive Feature Elimination (RFE) was done when predicting both Portuguese and Math final grades and the top 10 features were selected for both. For predicting Portuguese grades, the significant features in the subset of selected features were: address_U (1 if the student's home address is in an Urban area, 0 otherwise), Mjob_teacher (1 if the student's mother is a teacher, 0 otherwise), reason_reputation (reason for choosing this school is reputation), paid_por_yes (extra paid classes within Portuguese) and failures (number of classes student has failed in the past). When using the top 10 features from RFE, the logistic regression predicting Portuguese grade has an accuracy of 77.92%. Statistical Parity and Equal Opportunity were also calculated using this feature-set.

The same process was repeated for predicting Math. Of the selected features for predicting Math, the significant features were: sex, Mjob_services (mother works a civil service job), traveltime (home to school travel time) and failures (number of classes student has failed in the past). When using the top 10 features from RFE, the logistic regression predicting Math grade has an accuracy of 63.63%. Statistical Parity and Equal Opportunity were also calculated using this feature-set. All results from the analysis done with Feature Selection can be seen in Table 5.

5 Mitigating Bias

5.1 Removing the Protected Attribute

The first explored method to mitigate bias in this dataset was to remove sex, the protected feature. Then, using the features selected in RFE from the previous section, the same logistic regression analysis was performed to predict both Portuguese and Math final grades.

5.2 Augmenting the Training Dataset

The second method to mitigate bias in this dataset was to aument the dataset, done by duplicating the observations and switching the value of sex. This method is explained in Sharma et. al. [9]. The same logistic regression analysis is then conducted to predict both Portuguese and Math grades using the selected feature-set from the previous section. Results from both this and the previous method can be seen in Table 6, with the original RFE results included for comparison.

Model	Accuracy	Statistical Parity	Equal Opportunity
Portuguese (RFE)	77.92%	-0.172	-0.118
Math (RFE)	63.63%	0.203	0.190
Portuguese (sex removed)	77.92%	-0.117	-0.118
Math (sex removed)	62.34%	0.057	0.065
Portuguese (augmented)	76.62%	-0.096	-0.094
Math (augmented)	62.34%	0.057	0.065

Table 6: Mitigating Bias - Model Accuracy and Fairness Metrics

6 Discussion

It is interesting to see just from preliminary analysis, where trends start to appear in this dataset. When looking at grades for Portuguese, a language class, female students on average score higher than males, whereas males on average score higher than females in Math. Older students tend to score higher than younger students in both classes. The trends with respect to sex is particularly interesting given that there is strong societal associations with females to liberal arts and males to maths and sciences, so much so that there is even a published test from Harvard University called the Implicit Association Test, that reveals whether the test taker has such implicit bias (`https://implicit.harvard.edu/implicit/selectatest.html`). When looking at the p-values from the t-tests, we can see that there is a significant difference in grades between males and females in their final grades for Portuguese. This is also the case for age and Math; there is a significant difference in final math grades between older and younger students when looking at Table 4.

When applying methods to mitigate fairness in the dataset, there are trends that are consistent with what is known about these methods. When dropping the protected feature, statistical parity and equal opportunity both draw closer to zero, which is desirable as this indicates a smaller difference in probabilities of the two groups (male and female students) having a high predicted final grade for both Math and Portuguese. Interestingly, the accuracy in the models with sex removed do not drop for both Math and Portuguese. However, since the fairness metrics still improve, this could indicate that the models are not so dependent on Sex as a predictor for final grade despite our initial impression of the significant t-test results in the Preliminary Analysis. It also indicates that sex is not highly correlated with the final grade outcome variables, so it appears that simply removing the protected feature does perform well in mitigating bias while not having to sacrifice model accuracy.

There are more interesting results with the second method of mitigating bias, augmenting the dataset. In these results, the model accuracies for predicting both Portuguese and Math drop by small degrees (about 1%). Statistical Parity and Equal Opportunity draws even closer to zero for Portuguese when compared against the RFE models, while for Math there is no improvement in fairness metrics. This is consistent with what was observed in the exploratory analysis. The method of augmenting the dataset using synthetic data in effect "balances" the dataset by switching the values of the protected feature. In Tables 2 and 3, it is clear that the tabulated values are more equal and balanced with respect to final Math grades in Table 3, whereas the tabular values are less balanced for final Portuguese grades in 2. Given this context, it would be consistent that using synthetic data has more of a visible effect when predicting Portuguese grades than on the models predicting Math grades.

7 Conclusion

Following the example application from the Introduction of this project, if a dataset such as the one analyzed in this project were used to base decisions for a GATE program, students could potentially

have fallen victim to a biased model. Especially with regard to gender and STEM, it is not difficult to detect bias in datasets. There is common discussion of the implicit bias that exists where women are associated with social fields in academia such as history, writing, and english, whereas only men are strongly associted with STEM fields such as math, engineering, computer science, and so on. The goal of this project was to apply exploration and mitigation methods regarding fairness in a dataset with this in mind, and analyze the results when building prediction models off of such a dataset.

There were also limitations to this analysis. Firstly, the sample size was quite small, with only around 380 students. Ideally, a larger sample size would have been used to have more robust results. Another aspect is that students that do well in one subject probably tend to do well in other subjects, so high performing students have high grades across the board. Another aspect to consider is that there is no way to control for teachers. There is nothing included in the dataset regarding which or how many teachers the students had. While the last 2 aspects may be a good thing in some ways as it would introduce some noise and randomness in the data, given the small sample size, this aspect may have interfered with some results in unknown ways.

In a way, the implicit gender-STEM bias was confirmed through this analysis, emphasizing the importance and need for methods to mitigate bias such as the ones implemented in this project. Other ways that this study could be improved would be to include more students, more schools, and more subjects, all ways to increase sample size. Another interesting analysis would be to replicate these explorations but with linear regression - instead of binning final grades into a binary variable, which inevitably loses some granularity and information in the outcome features, predicting the final grades actual numeric values and doing a fairness analysis on those models would be an interesting exploration.

8 Code

All code and data can be found on `https://github.com/danielle0730/dsci531`

References

[1] Mutuku Christopher and Kiilu Redempta. Influence of demographic factors on academic performance among primary teacher trainees - a case study of machakos teachers college. *International Journal of Educational Studies*, 3(1):07–11, 2016.

[2] Paulo Cortez and Alice Silva. Using data mining to predict secondary school student performance. *EUROSIS*, 01 2008.

[3] Linda Green and Gul Celkan. Student demographic characteristics and how they relate to student achievement. *Procedia - Social and Behavioral Sciences*, 15:341–345, 2011. 3rd World Conference on Educational Sciences - 2011.

[4] Mehdi Kaighobadi and Marcus Allen. Investigating academic success factors for undergraduate business students. *Decision Sciences Journal of Innovative Education*, 6:427 – 436, 07 2008.

[5] Sotiris Kotsiantis, Christos Pierrakeas, and P. Pintelas. Predicting students'performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18:411–426, 01 2004.

[6] Susan Li. Building a logistic regression in python, step by step. *Towards Data Science*, September 2017.

[7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. 2019.

[8] Maliha Nasir. Demographic characteristics as correlates of academic achievement of university students. *Academic Research International*, 2(2):400, 2012.

[9] Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. Data augmentation for discrimination prevention and bias disambiguation. page 358–364, 2020.

[10] Tamara Thiele, Alexander Singleton, Daniel Pope, and Debbi Stanistreet. Predicting students' academic performance based on school and socio-demographic characteristics. *Studies in Higher Education*, 41(8):1424–1446, 2016.