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 Project Domain and Goals

As mentioned before, in addition to the primary goal of this project of investigating possible bias in this 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 'Student Performance Data Set' comes from secondary education data of two Portuguese schools and contains 649 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 math scores. 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 implementation of these definitions of fairness, prediction models to predict final math scores will be constructed and to examine and compare accuracy and fairness.

Other potential areas of this project include converting the outcome variable (a numerical grade) into a binary variable (High, Low) and running logistic regression models, repeating the analysis plan mentioned above and compare results. Secondly, this data set contains another table for the same group of students but for their grades in a Portuguese language class. We also are interested in repeating this analysis for these grades as well, as it could be interesting to compare results between a STEM class and a non-STEM class for the same group of students, and determine how biases may stay the same or differ across classes.

Preprint. Under review.

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 [9]. 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 consisted of predicting Portuguese and Math grades for the students, while looking at the 2 different protected features. Thus an analysis plan was constructed: 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 were 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 Statistical Analysis and Exploration

Preliminary exploratory analysis was done as an initial look into the fairness of the dataset. Although final grades are binned into a binary variable as described in the previous section, for exploration both numeric and binary variable versions of the final grade feature are used. Mean final grades were calculated with respect to each class and protected feature. Grades for each class with respect to sex are described in Table 1. Similarly, final grade averages were calculated with respect to age, older and younger, and can be seen in Table 2.

Class	Females	Males
Portuguese	13.137	12.039
Math	10.059	11.059

Table 1: Mean final grades for classes with respect to Sex

Class	Younger	Older
Portuguese	11.444	12.625
Math	7.889	10.639

Table 2: Mean final grades for classes with respect to Age

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 Tables 3 and 4.

Class	Sex	Age
Portuguese	0.0007	0.224
Math	0.063	0.084

Table 3: T-Test p-values for class grades (numeric) with respect to protected features

Class	Sex	Age
Portuguese	0.0003	0.804
Math	0.074	0.047

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

4 Analysis Methods and Results

A full logistic regression analysis was done to establish a baseline performance before implementing any methods to potentially mitigate bias for predicting Portuguese and Math grades (binary). Using all features to predict Portuguese grade, the full logistic regression model has accuracy of 80.52%. Statistical parity and equal opportunity were also calculated with these predictions with respect to sex and age. Repeating the analysis for Math, using all features to predict Math grade, the full logistic regression model has accuracy of 66.23%, and again statistical parity and equal opportunity were calculated with these model predictions with respect to sex and age. Formulas for statistical parity and equal opportunity are based on metrics described in Mehrabi et. al. [6]. These results can all be seen in Table 5.

Class	Model Accuracy	Statistical Parity	Equal Opportunity
Portuguese (Sex)	80.52%	-0.121	-0.020
Portuguese (Age)	80.52%	-0.404	-0.481
Math (Sex)	66.23%	-0.015	0.035
Math (Age)	66.23%	-0.603	-0.806

Table 5: Model Accuracy and Fairness Metrics for Portuguese and Math w.r.t. Sex and Age

5 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 predicting class grades, the logistic regression model performed better in predicting the Portuguese final grade than the Math final grade. Statistical parity metrics seem to range the most with respect to the 2 protected features when predicting Math grades - statistical parity is closest to 0 when the protected feature is sex, and statistical parity is furthest from 0 when the protected feature is age. The equal opportunity metric is also furthest from 0 when predicted math grades with protected feature age.

It appears that when predicting Portuguese grades, both statistical parity and equal opportunity stay close to 0, which is desirable and appears to be the most fair since this means the differences in the two groups, male and female, when predicting grades, is not biased. These metrics jump to around -0.4 for both values when the protected feature is changed to age. This trend is also consistent with predicting math - the metrics stay closer to 0 when the protected feature is sex, and gets further from 0 when the protected feature is age. This could mean that there is some lack of fairness coming from the teacher giving the grade when they take age of the student into account, perhaps unfairly giving older students higher grades from some preconceived notions that they are more mature or smarter. Or, this could simply mean that older students do in fact score higher in all classes because of more mental maturity, and better mental capabilities to perform better in school, or perhaps since they are closer in age to applying to university and higher education, they take their academics more seriously.

6 Next Steps

Next steps will involve exploring ways to mitigate bias such as removing the protected attribute and augmenting the dataset [8] and repeating the regression analysis and fairness metrics to evaluate for any improvement in mitigating bias. Once methods for mitigating bias are implemented, results will be further discussed to compare model accuracies and fairness metrics across all models and parameters.

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

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

[8] 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.

[9] 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.