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.' 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.

References

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[2] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54(6), 1–35 (2021)

[3] Sharma, S., Zhang, Y., R ıos Aliaga, J.M., Bouneffouf, D., Muthusamy, V., Varshney, K.R.: Data augmentation for discrimination prevention and bias disambiguation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 358–364 (2020)