Fairness and Privacy Assessment in Higher Education Student Performance Prediction

This project aims to investigate the fairness and privacy implications in the context of predicting higher education student performance based on a given dataset. The dataset contains features such as student age, gender, high-school graduation status, scholarship type, additional work, artistic or sports activities, partner status, total salary, transportation, and accommodation.

How we collect the data?

We are using the “Higher Education Student Performance Prediction” dataset, which is publicly available from the UCI Machine Learning Repository. This dataset has been compiled and made accessible for research and analysis purposes. It contains information related to higher education students and various attributes that are relevant to their academic performance, as well as other personal factors. These features provide valuable insights into the students' demographic, socio-economic, and lifestyle characteristics.

What kind of statistical study we wish to perform?

We can consider different statistical studies. For example, descriptive statistics can be used to summarize and understand the dataset. Calculate measures such as mean, median, standard deviation, and quartiles for numerical variables like student age, total salary, and transportation costs. For categorical variables like gender, high-school graduation status, or scholarship type, calculate frequencies and percentages.

Then we can examine the relationships between different variables. For example, you might investigate the correlation between student age and academic performance or explore whether scholarship type is correlated with additional work. Moreover, we can perform hypothesis testing to determine whether there are statistically significant differences between groups. For example, we can test whether there is a significant difference in academic performance between students who graduated high school and those who didn't.

What possible privacy problems that we will identify and address?

Privacy issues are paramount when dealing with data related to individuals, especially in educational and personal contexts. In this dataset, which includes features such as partner status, total salary, and other personal attributes.

The dataset contains sensitive personal information that, if not properly protected, can lead to the identification of individuals. Even if the dataset is anonymized, there is a risk that determined individuals could be re-identified. This can happen through the combination of different attributes or by cross-referencing the dataset with external data sources. Combining this dataset with other publicly available datasets may increase the risk of re-identification or unintentional data leakage. Adversaries might

use the available information to infer sensitive details about individuals, such as inferring a student's income or personal life based on partner status and other attributes. Some attributes in the dataset might reveal sensitive information about an individual's personal life, which could be exploited by malicious actors.
To address these privacy issues, consider implementing privacy-preserving techniques such as Randomised Response, Laplace Mechanism and Exponential Mechanism. Adding noise to query results to protect individual data while still providing useful aggregate information.

What issues with fairness that we want to measure and address?
For this dataset, we will encounter both Group and Individual Fairness problems.
Analyzing and modeling sensitive attributes like gender can lead to biased outcomes or perpetuate existing biases. In order to address this issue, we can implement fairness-aware machine learning models and fairness metrics to detect and mitigate bias in the analysis. Ensure that gender or other sensitive attributes do not lead to unfair outcomes.
An imbalance in the dataset can lead to the underrepresentation of certain groups, affecting the model's ability to make fair predictions. This can be handled through techniques such as oversampling the minority category, undersampling the majority category, or generating synthetic data while considering fairness constraints.
Those are about Group Fairness, and we also run into Individual Fairness problems in this dataset. Because the final scores are discrete - the scores are categorised as 1-7, meaning that people who may have performed the same way, get scores of the difference could be greater than or equal to 1 - which is not reasonable. We might consider defining a measure of "similarity" and then ensuring that individuals with similar measures get similar predictions.

About Reproducibility.
1. The dataset we use is publicly available and anyone can use the exact same dataset as us;
2. We use a combination of colab and github to share our code, making it possible for anyone who wants to use our methods to get their hands on the exact same code as us;
3. For stochastic processes, such as Randomised Response, we record the random seed we have used to ensure that someone else can get the same result using that random seed;
4. We will also divide our dataset into a training set and a test set, and use cross-validation to ensure that the results on the test set are the same as the results on the training set, after several tests;
5. We will also use bootstrapping to get estimates of uncertainty in Python.