Assignment 3: Fairness

Oliver Miller (260911415)
McGill University Department of Electrical Engineering

1 Fairness concerns

1.1 Data access

Hospital staff should be given access to all information used by Prioritize. Hospitals are constantly in possession of sensitive information, thus giving them access to the information used by TriageAssist will not lead to any concerns. Patients should be given access to all information on their case (and not information regarding other patients found within TriageAssist, so as to avoid any conflicts between patients).

I believe the decision of who gets access to these resources should be a joint effort between the upper employees at Prioritize and the hospital administration staff. The combined knowledge of these two groups will yield the most effective distribution of information possible.

How they go about deciding who gets what information should be decided through a simple discussion between the two parties. I think that both parties have the interest of the patients as their core values, meaning they will come to a conclusion that benefits as many people as possible.

1.2 Privileged groups and favoured outcomes

In my code, I explore two combinations of privileged and unprivileged groups. The first is sex: male versus female. The second is race: White versus Black, Asian, Hispanic and other. These definitions or privileged and unprivileged are based solely on societal stereotypes that exist in our world. However, this does not consider the medical difference between the groups (i.e., there exist cases where differences in sex or race have a medical implication on the status of the patient). A favoured outcome, regardless of the identified group, would be a more urgent status as define by TriageAssist, as this means the given patient will be treated earlier than other, less urgent patients. An alternative interpretation of the favoured outcome would be lower urgency, as this would mean a less critical condition of the patient which could be considered favourable with regard to the health status of the given patient.

1.3 Fairness concerns

One potential fairness concern is prioritizing patients based on attributes that should have little to no effect on the results of the TriageAssist platform, such as sex or age, beyond their medical importance. Further fairness concerns may arise should any of the Prioritize team responsible for the development of TriageAssist have any personal investment in the outcomes assigned by their model. If this is the case, certain outcomes for certain individuals may be considered more favourable from the perspective of the Prioritize employee.

2 Fairness metrics

Below are the calculated fairness metrics for the training data:

Fairness Metric	**Male/Female**	**White/Black, Asian, Hispanic, other**
Statistical Parity Difference	-0.3594907407407408	-0.05821676078028748
Disparate Impact	0.4190048634493079	0.900643394934976

Further, for the testing and predicted data, the fairness metrics are:

Fairness Metric	**Male/Female**	**White/Black, Asian, Hispanic, other**
Statistical Parity Difference	-0.5101337086558762	-0.125
Disparate Impact	0.23323460968902054	0.8125
Equal Opportunity Difference	-0.5254658385093167	-0.09196025293586263

In statistical parity, fairness means people in both privileged and unprivileged groups have equal probability of being assigned the positive predicted class. In mathematical form, this can be defined as:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1) \tag{1}$$

where A is some protected attribute. The statistical parity difference is simply the subtraction of both sides of the above equation. As such, in order to be the most fair, the statistical parity difference should be as close to 0 as possible, where 0 would mean a perfectly fair scenario. With this in mind, from the above data, the race group has a better statistical parity difference than the sex group. It is important to note that the race data is synthetic, meaning we can discuss it's implications in the context of this assignment, though it has no bearing on the actual dataset or its applications.

Disparate impact is measured as the ratio of the rate of a positive outcome for a disfavored group to the rate of positive outcome for the favored group. In an ideally fair scenario, disparate impact should be equal to 1. As such, with regard to the above data, once again the race group is seemingly more fair according to the disparate impact fairness metric.

In equal opportunity impact, fairness means both privileged and unprivileged groups have equal true positive rate, or they get the positive outcome at equal rates. Mathematically, this is defined as:

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1) \tag{2}$$

where A is some protected attribute. The equal opportunity difference is simply the subtraction of both sides of the above equation. Similar to statistical parity difference, an ideally fair scenario would have a equal opportunity difference of 0. This means that the race group is more fair than the sex group with regard to equal opportunity difference.

Statistical parity difference and disparate impact would be suitable metrics by which to assess fairness. Both metrics have to do with the equalizing the positive outcomes assigned to both privileged and unpriv-

ileged groups. For Prioritize, this should be a top priority to avoid any and all prejudice when assigning outcomes through TriageAssist.

3 Pre-processing for fairness

The pre-processing technique used is Learning Fair Representations (LFR). This pre-processing technique yielded the following fairness metrics:

Privileged group/unprivileged group	**Statistical parity difference**	**Disparate impact**
Male/Female	0.0	1.0
White/Black, Asian, Hispanic, other	-0.0771143480492813	0.700889801505818

Great improvements can be seen in the male/female group as the ideal values of 0.0 and 1.0 have been calculated for statistical parity difference and disparate impact, respectively. Similar results are not observed for race groups. However, this can be attributed to the fact that this column is comprised of synthetic data, meaning there is not actual correlation between this feature and the label of the dataset.

LFR is a pre-processing technique which encodes the data and also obscures information about protected attributes, favouring individual fairness in the dataset. This is particularly applicable to this dataset as avoiding preconceptions given sensitive features is key for avoiding bias in the TriageAssist model. This holds obvious benefits including obscuring potentially bias-inducing information; however, especially in the case of a model being used in the medical field, obscuring information could potentially lead to misclassification which can be extremely harmful to certain patients.

Another potential pre-processing technique that could be used is relabeling. Relabeling involves identifying rows of unfair decisions and changing the outcome of said. Though this may prove more effective than LFR at improving fairness metrics, in the medical field, simply changing outcomes in favour of fairness is not an effective strategy given the severity of the decision-making. Medical outcomes are not ones to be messed with due to the potentially harmful implications on the patients health, thus labels cannot simply be changed in order to protect fairness.

In this particular case study, individual fairness should be prioritized, thus LFR is an ideal choice for a pre-processing technique.