ECSE 557

Intro to Ethics of Intelligent Systems

Assignment 3:

Fairness

CHIKHAOUI Hamza 260912960

Winter 2024

Part 1 : Fairness concerns :

1) a)

- The Prioritize engineering team / executive team have direct access on the dataset containing medical and personal information of patients, including the knowledge of them having (or not) a form of a heart disease.
- The nurses and the doctors would have access to the output of TriageAssist that will predict whether or not the new patients are at risk of a heart disease.
  Moreover, one could assume that the doctors and nurses would also have access to the dataset of medical and personal information of existing patients.
  One could also note that they would have similar information about new patients at the entry of the ER. Then, although it is not stated in the case study, one may argue that the nurses would use the newly collected data to populate the old dataset of patients, thereby giving the Prioritize team access to more data about the newly admitted patients.

b)

- The patients decide to give their consent to their information being collected in a dataset. One might however argue the extent to which their consent is full, as they often reach the hospital in dire need for health care, which incentivizes them to give their consent regardless of the consequences. Moreover, the patients do not necessarily know on the spot that their data is being recorded.
- It is evident that the owner of the dataset decides who to share the dataset with. Although it is not specifically specified, one might assume that the dataset is composed of samples from different hospitals, in which case the ownership of the dataset would be split between the various hospitals that provided data and the Prioritize team. If not, (i.e. the data is only coming from the Montreal hospital), then the ownership of the dataset would then be split between the Prioritize team and the hospital. One may also state that there could be a scenario in which the dataset is solely owned by the Montreal Hospital, and that the hospital only provides a restricted access to Prioritize.
- The Prioritize executive team decides who in the engineering team will have access to the dataset, and the extent of the access provided (they might for example give an access to a version of the dataset that does not contain direct identifiers about each sample).
- Finally, the hospitals executives select who in the health services team (which nurses, which doctors) will have access to TriageAssist, and who will have access to the historical data about the patients of the hospital.

c) In my opinion, the question of who decides to grant access to what information is the product of the following process:

- First of all, there is little information about the ownership of the dataset. However, in the case of a shared ownership between the hospital and Prioritize, both would be responsible for who has access to the data, and the extent to which they have access to the data. In the case where the hospital would be the sole owner of the dataset, then the hospital would have complete responsibility over who has access to the dataset. That would in turn imply that the hospital would be the only stakeholder that will decide who has access to the data.
- The hospital would base the decision of granting access or not to the data through the intent of protecting the patients as much as possible, while also protecting the health and privacy of their

health practitioners. This includes sharing the data in exchange for TriageAssist if that would imply to reduce the congestion of patients in the ER.

- In the case of the patients, the decision of accepting to share their data is made when requesting for urgent care at the ER, in the optic of speeding up the process of them being taken care of. Therefore, one could ask the question of whether or not they are really consensual in giving their data, or if they do so because they are in a position of weakness when getting to the ER.
- The Prioritize executive team decides who in the engineering team has access to the dataset, and decides who is allowed to use TriageAssist. As a company whose main focus is to be profitable, Prioritize will base its decisions on the desire for the company to generate the best product possible, and will also try to partner with the highest number of hospitals. This would in turn imply to share the dataset with the largest amount of competent people possible to optimize the fine tuning of TriageAssist.
- The nurses and the doctors decide who gets access to medical care based of the predictions provided by TriageAssist. Therefore, this means it is effectively TriageAssist which choses who gets access to urgent medical care.

2) From question 1, it seems evident that the patients of the hospital are at a disadvantage in this case study. They are the most compelled to provide their data.
When studying the data set, we observe a large disparity in the distribution of sex:

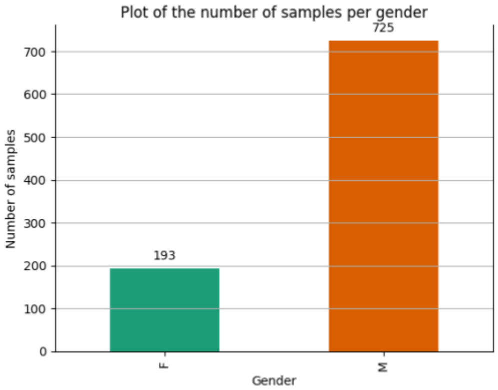


Figure 1 : Distribution of sex across our dataset

We can see that the dataset contains substantially more male samples (725) than female samples (193). This in turn means that women are disadvantaged as their predictions for heart disease are more likely to not be accurate, while the model will be primarily trained for heart disease detection in male samples. This is further shown in the following figure :

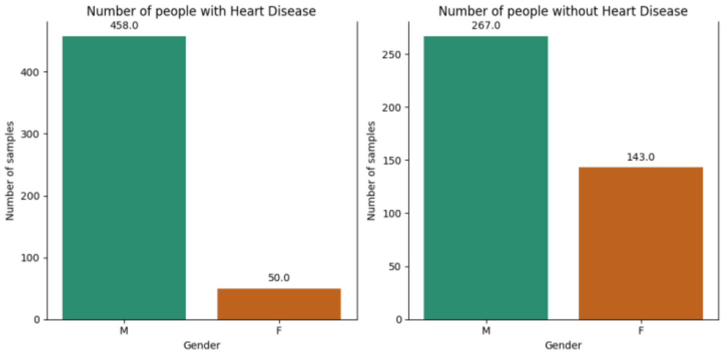


Figure 2 : Distribution of heart disease across genders

From Figure 2, we can see that there are about 9 times more male samples with heart disease then there are female samples with heart disease. However, there is "only" about 1.8 times more male samples without heart disease than there are female samples without heart disease. This highlights a clear imbalance in the dataset as this distribution would theoretically imply that the model would be more biased in detecting a heart disease in both genders than not detecting one. Considering that the main goal of the algorithm is to correctly detect the presence of a heart disease (so that urgent care can be provided efficiently), this would make it even less likely for the algorithm to be able to detect heart diseases in the unprivileged group (female), putting them further at risk.

Finally, one could also notice that the ratio of healthy subjects to ill subjects is much higher for women (about 3 to 1) than men (about 2 to 1). Thus, we conclude that the privileged group is men, and the unprivileged one is women.

We will also note that the dataset is much more imbalanced when it comes to gender distribution than race distribution as one can see from the following plot :

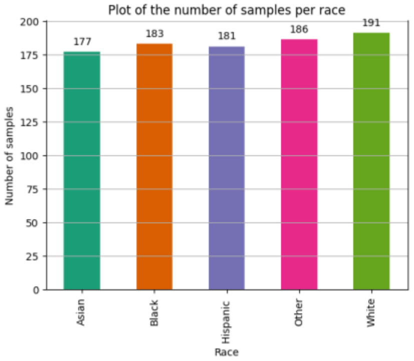


Figure 3.a : Distribution of race in our dataset

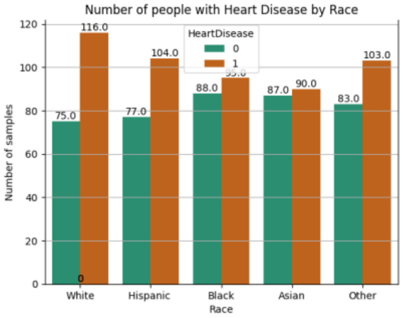


Figure 3.b : Distribution of heart disease across race

The presence and absence of the disease across all races is more homogeneously distributed.

The main goal of TriageAssist is to assist the Montreal hospital in de-congesting the ER. Therefore, the favored outcome of this assistive model is detecting which patient is at risk of a heart disease (i.e. a true positive result).

3) Based on the previous analysis, we can be concerned about how fair this model will be across genders. Our analysis showed that there are many more men than women in the dataset, and that these man are (statistically) more likely to have a heart disease then women. Therefore, men will in turn be more likely to be classified as having a heart disease, which implies that they will have a more favorable outcome (and would be receiving more urgent care than women).
Furthermore, we can expect the predictions to be more accurate for men than women because of the distribution of the heart disease within each genders, and because there are more men samples.

Part 2 : Fairness metrics:

4) Considering that the fairness metrics will be computed to assess the performance of the model and the original dataset, we will select the metrics from the ClassificationMetric class.
We obtained the following results :

Table 1 : Classification metric performance of our model

Metric	Performance
True Positive Rate difference between unprivileged and privileged groups	-0.269
False Negative Rate ratio between unprivileged and privileged groups	3.05
Average Odds Difference between unprivileged and privileged group	-0.123

A table breaking down the results from all metrics obtained is available in the appendix section of this report, alongside a short explanation of each metric and how they are computed. As per the selected metrics, they were chosen for the following reasons :
- True positive rate (TPR): measures the ability of the model to correctly identify positive instances. This is particularly relevant in our case (medical diagnoses), as it would direct us to provide specific care to positive samples.
- False negative rate (FNR): measures the proportion of actual positive instances that are incorrectly classified as negative. This metric is particularly relevant in situations where the cost of missing positive instances is high, which is our case as mis-diagnosing someone may in turn induce that this person would not receive critical care where they should.
- Average odds difference (AOD) : This metric considers both false positive rate and true positive rate, providing a balanced view of disparities in both directions. This metric in turn helps us assess overall balance in error rates in both groups.

5) We observe that from the achieved results, our logistic regression model suffers from some fairness flaws. We start off by noticing that negative values correspond to a smaller value of the metric for the unprivileged group than for the privileged group.
We see that the True Positive Rate difference (TPR difference) is negative. Therefore, the TPR is lower for female samples than male samples, which indicates that our model is not fair in detecting a heart disease in both groups. It essentially does better in detecting heart diseases for male samples rather than female samples. This value should ideally be 0, in which case a heart disease is as easily detected in a men and in a women.
Moreover, the False Negative Rate (FNR) ratio is quite large and positive (exact value of 3.05). The FNR ratio is defined as being the ratio of the FNR of the unprivileged group over the FNR of the privileged group. This in turn means that women are 3 times more likely to be given a false negative rate then men. Ideally, this value should be 1 in a perfectly fair scenario.
Finally, the average odds difference between female and male is obtained to be -0.123. The average odds difference is defined as the average of difference in FPR and TPR for unprivileged and privileged groups, which provides an estimate for the equality of odds in both populations. This means that men are more likely to get classified as having a heart disease (and receive special care) than women, which is unfair to the women.

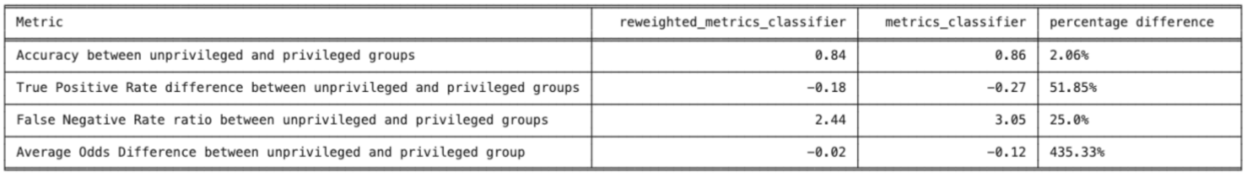
6) The outcome of our model is both assistive, in the sense that our model aims to provide assistance to patients that are predicted to have a heart disease. On the other hand, a false negative prediction of our model would be punitive to the individual that would not be receiving the care they need. Therefore, when classifying whether people have a heart disease or not, it is imperative for our model to have a low rate of false negatives. This should be achieved even if this leads to a high rate of false positives. Indeed, the cost of not predicting a heart disease in a patient suffering from a heart disease is much higher than the one associated with predicting a heart disease to a patient that does not have one. For this reason, it is important that both the privileged and the unprivileged have similar values of FNR. Therefore, the FNR ratio should be as close to 1 as possible, thereby guaranteeing fairness with that regard for both groups.
In the same spirit, we are interested in providing the most care to the largest amount of people who need it (without classifying everyone as positive). Therefore, the average odds difference is also an acute metric of fairness in our context. In our context, we aim for an average odds difference of 0 between male and female, which would imply that men and women are classified as having a heart disease at the same rate (irrespective of whether they have a heart disease or not).
Finally, the cost of a false positive prediction is null for the patient, but may have some implications for the hospital. One can argue that a false positive prediction would allocate too many resources to a patient that does not require them, while those same resources could have been used to help a patient who actually suffers from a heart disease.
It is also important to maximize the true positive rate to make sure that (at least) all the patients that suffer from a heart disease are receiving the care they need.

Part 3: Pre-processing for fairness

7) The pre-processing mitigation technique that I have decided to implement is the reweighting technique. Reweighing is a preprocessing technique that weights the examples in each (group, label)

combination differently to ensure fairness before classification. The implementation of this technique is available in the Jupiter notebook submitted alongside this report.
The following results were achieved :

Table 2 : Summary of the results obtained for the regular and reweighted classifier.



Metric	reweighted_metrics_classifier	metrics_classifier	percentage difference
Accuracy between unprivileged and privileged groups	0.84	0.86	2.06%
True Positive Rate difference between unprivileged and privileged groups	-0.18	-0.27	51.85%
False Negative Rate ratio between unprivileged and privileged groups	2.44	3.05	25.0%
Average Odds Difference between unprivileged and privileged group	-0.02	-0.12	435.33%

8) After reweighting the data, we see a notable improvement in all the metrics that were previously computed.
First, we observe that the TPR difference was reduced to about 51.85% of the original value. This means that women have a TPR that is closer to that of men, which in turn means that women are more likely to be given a true positive (or men are made less likely to get a true positive).
Furthermore, the false negative ratio was reduced by 25% which implies that now women are “only” 2.4 times more likely to receive a false negative than men. Naturally, this value remains substantially high (even after the 25% reduction due to reweighting), and further work should be conducted to reduce this value even more.
Finally, the metric onto which reweighting had the largest effect on is the average odds difference (AOD) between women and men, which is a crucial benefit to our model. Indeed, the AOD was reduced by about 435%, to a value close to 0 (-0.02). This means that after reweighting, men and women have very similar true positive and false positive rates. Indeed, the AOD is defined as the average of the difference in FPR and TPR for the unprivileged and privileged groups. Therefore, that means that not only are women more accurately classified, but even erroneous classifications are giving them a more favorable outcome. This re-equilibrates the fact that women and men who have heart diseases are being taken care of at the same satisfactory rate, but also that men and women who do not have heart diseases are erroneously being taken care of at similar rates, which is more fair.
One shortcoming of the implemented method is that we noticed a slight loss of accuracy of our model. Although still acceptable, one would have to confront this loss of accuracy with how crucial it is for our model to correctly identify patients who might suffer from heart diseases.
One preprocessing method that was not implemented is the disparate impact remover. The disparate impact remover is a preprocessing technique that edits feature values increase group fairness while preserving rank-ordering within groups. It is reasonable to expect a disparate impact between men and women as there is most likely a different prevalence of a heart disease in their respective demographic groups. For example, if about 10% of men suffer from a heart disease, but only 5% of women suffer from a heart disease, our model should be able to account for a disparate impact as men are most likely to suffer from a heart disease then women, in which case our model can benefit from such a pre-processing technique. However, the implementation of this method might not be able to reduce the AOD to the extent to which reweighting did. In our case, the most important fairness consideration is that men and women receive the same amount of care, whether they respectively suffer from a heart disease of not. That would imply the reduction of our model’s bias towards men, even if more men are to be classified as having a heart disease than women.

In summary, we can summarize the pros and cons of both methods (reweighting and disparate-impact remover) in the following table :

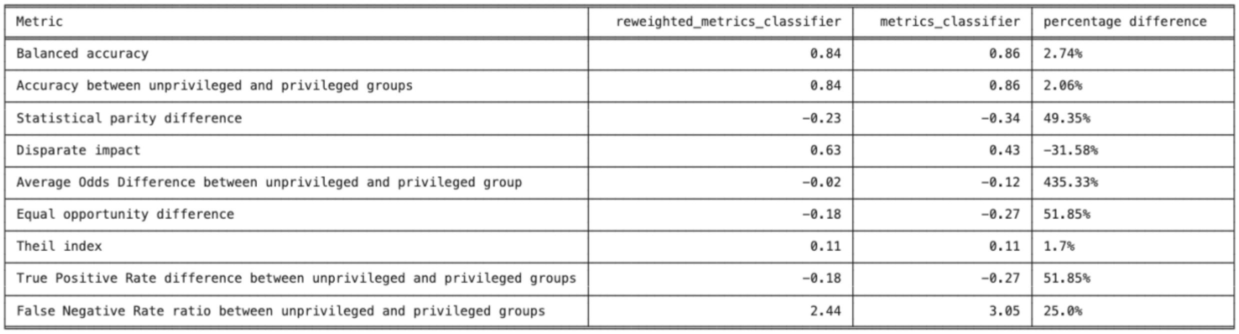
Table 3 : Summary of the Pros and Cons of Reweighting and Disparate-impact remover based on material covered in class

Method	Pros	Cons
Reweighting	- This method allowed for a decrease in the TPR difference, the FNR ratio, and the average odds difference between the privileged and unprivileged group. - Another advantage of this pre-processing method is that it has done so without modifying any value, feature or label of our dataset.	- One drawback of this method however is that this was done at the cost of some accuracy of our classifier. - As seen in class, this method only worked because all possible combinations of privileged (p) to unprivileged group (up) and favorable (f) to unfavorable (uf) outcomes were present in the dataset. This method would not have worked if any of the following combinations (uf-p, uf-up, f-p, f-up) was not present in our dataset.
Disparate-Impact Remover	- Preserves the rank ordering within our groups	- Involves editing feature value internally

Appendix

Full table of all the metrics obtained before and after reweighing. Note that some indicate the same values (such as the Equal Opportunity Difference and the True Positive Rate difference).

Table 4: Summary of all metrics calculated before and after reweighting :



Metric	reweighted_metrics_classifier	metrics_classifier	percentage difference
Balanced accuracy	0.84	0.86	2.74%
Accuracy between unprivileged and privileged groups	0.84	0.86	2.06%
Statistical parity difference	-0.23	-0.34	49.35%
Disparate impact	0.63	0.43	-31.58%
Average Odds Difference between unprivileged and privileged group	-0.02	-0.12	435.33%
Equal opportunity difference	-0.18	-0.27	51.85%
Theil index	0.11	0.11	1.7%
True Positive Rate difference between unprivileged and privileged groups	-0.18	-0.27	51.85%
False Negative Rate ratio between unprivileged and privileged groups	2.44	3.05	25.0%

Description of each metric used in this report :

Certainly, here's the text formatted for easy copying into Word:

1. **True Negative Rate (TNR):** The proportion of actual negatives correctly predicted as negatives; TNR = TN/(TN+FP).

2. **Accuracy between Unprivileged and Privileged Groups:** The overall accuracy of the model, measuring the ratio of correct predictions to the total predictions.

3. **Statistical Parity Difference (SPD**): The difference in favorable outcomes between unprivileged and privileged groups; SPD = P(prediction|unprivileged) - P(prediction|privileged).

4. **Disparate Impact (DI):** Measures the ratio of favorable outcomes between unprivileged and privileged groups; DI = P(prediction|unprivileged) / P(prediction|privileged).

5. **Average Odds Difference (AOD):** The average difference in true positive rates and false positive rates between unprivileged and privileged groups.

6. **Equal Opportunity Difference (EOD):** The difference in true positive rates between unprivileged and privileged groups.

7. **Theil Index:** Measures the inequality in prediction distribution; Theil = (1/N)Σ(i=1 to N){(pi/μ)ln(pi/μ) + (1-pi)/(1-μ)ln((1-pi)/(1-μ))}.

8. **True Positive Rate Difference:** The difference in true positive rates between unprivileged and privileged groups.

9. **False Negative Rate Ratio:** The ratio of false negative rates between unprivileged and privileged groups; FNR_ratio = FNR_unprivileged / FNR_privileged