Assignment 2

Part 2

ECSE 557 - Introduction to Ethics of Intelligent Systems

Winter 2022, McGill University

Gauri Sharma
261026894

Question 1

Who gets access to what resources/information in this case study?

- The company and the designers working on TriageAssist will get access to the patients' information and they might also get access to the hospitals' information from where they tried to get the patient's data.
- The hospital and the doctors and nurses who are working there will have access to all of the patients' information.
- If the patient consent is taken for the development of TriageAssist using his data, then patient will have the information regarding what the technology is and (maybe) how it is being developed.
- The doctors, nurses and patients get access to the technology and methodology being employed in building of that technology. How their data is being used? For what purpose it is being used? These are some pieces of the information they should get access to from Prioritize.

Who decides who gets access to these resources?

- The patient whose data is being shared with the company should decide if they want to share their data with the company for TriageAssist or not.
- The hospital also decides what all patient data and hospital data they are sharing with Prioritize for TriageAssist.
- The doctors and the nurses who are in direct contact with the patient and can comment better on their health can also decide what all information needs to be shared.
- Prioritize will decide what all information they will provide about TriageAssist to the patients, doctors and nurses.
- The designers working on the TriageAssist will decide what all information they give to Prioritize regarding the technology being developed.

How do they decide who gets what?

- The doctors and nurses will decide based on the needs of TriageAssist and privacy of patients what all information they can provide to Prioritize.
- The patient will decide based on their privacy what information should be provided. The idea of trust will also decide if they want to share their information with Prioritize or the hospital. Second to that, they can provide information based on the needs of TriageAssist and their own interest.
- The hospital will decide based on their own privacy and legal constraints if they can provide this information to Prioritize or not.
- Prioritize will decide based on the idea of transparency and privacy whether they should share their and TriageAssist's information with the hospital and patients or not.

Question 2

- **Privileged Group(s)** - Age (less than 38), Sex (M), Race (White), ChestPainType (ATA), ExerciseAngina (N)
- **Favoured Outcomes** - HeartDisease (1), ExerciseAngina (Y)

Question 3

- There are possibilities that a particular race might get assistance in a better and quicker form, from TriageAssist.
- People of certain age group are likely to be treated with better treatment options. TriageAssist might not provide efficient treatment options to younger generation assuming they are fit.
- Equitable finances is another issue that might arise. People who are paying treatment bills at time or can afford it might get assistance quicker.
- Patient and provider autonomy needs to be considered as well, we need to see if the patient wants to be assisted by TriageAssist or not. Even the doctors based on their personal biases can direct a patient to TriageAssist.

Question 4

Code block can be found in *Assignment2_Part2.ipynb*

Question 5

Fairness metrics			
Fairness Metric	**Train Value**	**Test Value**	**Implication**
Statistical Parity Difference	-0.193351	-0.223776	This tells us that the difference in mean outcomes between unprivileged and privileged groups. It basically tells that the members of each group have the same chance of receiving a favourable output.
Disparate Impact	0.685805	0.664336	This is the disparate impact outcomes between unprivileged and privileged groups. It compares the proportion of individuals that receive a positive output for the two groups we have.
Smoothed Empirical Differential Fairness	0.401399	0.497380	This is the smoothed empirical differential fairness outcome between unprivileged and privileged groups. Smmothed EDF tells us that regardless of the combination of protected attributes, the probabilities of the outcomes will be similar.

Question 6

- **Consistency** - It is an individual fairness metric that measures how similar the labels are for similar instances. It will help us understand how consistent the technology is for similar set of conditions.
- **Confusion Matrix** - This computes the number of true/false positives/negatives, optionally conditioned on protected attributes. This will be helpful in understanding the effect of various protected attributes on the technology.
- **TPR Difference & FPR Difference** - This will return the difference between the true positive rates/ false positive rates of unprivileged and privileged groups. This will help us understand if the technology is giving correct results a particular group or not and is important since it being employed in the healthcare domain.
- **Error Rate Difference** - This is the difference in error rates for unprivileged and privileged groups. This will help us understand if the technology is biased towards a particular group or not.

Question 7

Code block can be found in *Assignment2_Part2.ipynb*

Question 8

Reweighing technique was used in pre-processing for the mitigation of bias in the dataset. This technique weights the examples in each (group, label) combination differently to ensure fairness before classification. Reweighing affected the fairness metrics, it decreased the statistical parity which ensured that our fairness is maintained. Similarly, our disparate impact also reduced a bit, which again showed that the bias has been mitigated.

Benefits/Shortcomings/Appropriateness

This technique is useful as by carefully choosing weights, the dataset can be made discrimination free, without having to change any labels. The weights can be used directly in any method based on frequency counts. It does fall behind *massaging*, another technique to mitigate bias, but still gives reasonable performance.
Here it was useful as we had a significant amount of data and that was supposed to be used for a classification task.

Reweighing vs LFR

Learning fair representations (LFR) is a pre-processing technique that finds a latent representation which further encodes the data well but it makes the information about protected attributes unclear, which is a trade-off if we compare it with re-weighing because no information is being left obfuscated in case of reweighing.

References

Christoph Baur, Shadi Albarqouni, and Nassir Navab. 2018. Generating highly realistic images of skin lesions with gans. *CoRR*, abs/1809.01410.

Nicholas Caporusso. 2021. Deepfakes for the good: A beneficial application of contentious artificial intelligence technology. *Current Issues in Tourism*.

Chen T.Y. et al Chen R.J., Lu M.Y. 2021. Deepfake: a social construction of technology perspective. *Synthetic data in machine learning for medicine and healthcare*, 5:493–497.

S M Abrar Kabir Chowdhury and Jahanara Islam Lubna. 2020. Review on deep fake: A looming technological threat. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. 2018. Gan-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *CoRR*, abs/1803.01229.

Andrei O. J. Kwok & Sharon G. M. Koh. 2021. Deepfake: a social construction of technology perspective. *Current Issues in Tourism*.

Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 2019. Deep learning for deepfakes creation and detection. *CoRR*, abs/1909.11573.

ORI. 2019. Foresight into ai ethics (faie): A toolkit for creating an ethics roadmap for your ai project.

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine Andriole, and Mark Michalski. 2018. Medical image synthesis for data augmentation and anonymization using generative adversarial networks.

Jessica Silbey and Woodrow Hartzog. 2019. The upside of deep fakes. *Maryland Law Review*, 78.