Assignment 3

ECSE 557 - Introduction to Ethics of Intelligent Systems

Winter 2022, McGill University

Gauri Sharma
261026894

Part 1

Question 1

Complete the AIA questionnaire.
The AIA questionnaire was answered and is attached under the file name ”aia-en.pdf” and ”aia-en_new.pdf” (this contains the score after changing mitigation strategies).

Question 2

Summarize the results obtained from the AIA and what mitigation strategies you considered to reduce the risk level of TriageAssist.

The AIA assessment gave the following results

- Impact Level : 2
- Current Score: 53
- Raw Impact Score: 53
- Mitigation Score: 27

An impact level of 2 implies that the system will have a moderate impact and the score range is from 26% to 50%

Broadly, the following mitigation strategies were considered. (Detailed description can be found on page 5 of the ”aia-en.pdf”):

- Considered various internal and external stakeholders.
- Documented the processes to test datasets against biases and other unexpected outcomes and for resolution of data quality issues
- Made certain information publicly available.
- Did an audit trail that provided relevant information regarding the decisions being made and the system log (example -recording recommendations made by system, identifiable decision points, up-to-date log of decisions and changes made to the system)
- Prepared a concept case (Montreal hospital pilot study) to the Government of Canada Enterprise Architecture Review Board
- Designed and built security and privacy into systems from the concept stage of the project

Question 3

Create a .pdf report of your AIA once you complete the questionnaire. You will submit this as part of your deliverables.
The pdf file was created and is attached under the file name "aia-en.pdf" and "aia-en_new.pdf" (this contains the score after changing mitigation strategies).

Question 4

How does the impact score change when you remove one of the mitigation measures you've outlined in Question 2? Considering this, reflect on how much impact single mitigation strategy can have on the overall impact of an AI system.
The impact score didn't change for me when I changed one or more than one mitigation techniques. However, when I tried to fill in a "No" for all the answers in the mitigation techniques, the impact score fell down to 45, although the raw impact score stayed at 53. Therefore, according to my observation, changing single mitigation technique won't have a significant impact on the impact score and thus, the AI system.

Part 2

Question 2

Model card elements:

1. **A brief description of the model**
   The logistic regression model was used to predict the presence of heart disease in an individual. It does not aim to detect the cause of the disease. The values predicted were compared with the true values and based on that fairness metrics were calculated. The model was standardized in a way where the values were distributed such that the mean of observed values is 0 and the standard deviation is 1. The model was reweighed to mitigate any bias that might be arising and thus, improve the fairness metrics.

2. **Input data**
   The input data was mostly numeric, which contained readings for various medical tests (example: heart rate, blood pressure etc). It also consisted of information regarding the age and gender of the person.

3. **Output from model**
   The model predicted the heart disease in the patient. We got the following outputs for every model run:

   - Prediction of heart disease.
   - Performance metrics of non DP and DP classifiers.
   - Values of the following fairness metrics: Difference in mean, disparate impact and smoothed empirical differential fairness outcomes.
   - The fairness metrics were again obtained after reweighing the model.

4. **Model architecture**
   Logistic regression with StandardScaler standardizing technique was used. AIF360 library from IBM was used for the calculation of fairness metrics and mitigation of bias in the model.

5. **Performance of the model illustrated through two different metrics (you can use the metrics you worked with in Assignment 2). For example you can illustrate performance of the model via use of accuracy and recall.**
   The model performance was obtained based on the metrics given in figure 1 for original dataset and figure 2 for the differential private classifier.

- The confusion matrix for respective classifiers is shown in figure 3 and 4. It shows that the true positive values for normal classifier were 95 (Refer 1,1 on plot) whereas true negative were 59 (Refer 0,0 on plot). It means classifier correctly predicted these many values.
- The value 0 indicates the person doesn't have heart disease wherease 1 indicates it has heart disease.
- The accuracy was calculated on the test dataset and the normal classifier performed better by giving a score of 83.7%, whereas DP classifier gave 73.37%,
- Overall, the non-DP classifier performed better. Support is the measure of how frequently an item appears in the dataset.
- As seen, the precision is higher in non-DP classifier (83%) which indicates that it returns more relevant results as compared to DP.
- High recall of 77% in non-DP classifier indicates it returns "most" of the relavant results whether or not irrelevant ones are also returned. This is comparatively lower for DP classifier.
- F1 score is the measure of how perfect, precision and recall are. As observed precision and recall for non-DP are near to perfect with a score of .80 out of 1.

```
Non-private test accuracy: 83.70%
              precision    recall  f1-score   support

           0       0.83      0.77      0.80        77
           1       0.84      0.89      0.86       107
```

Figure 1: Predictions from original testing data

```
Private test accuracy: 73.37%
              precision    recall  f1-score   support

           0       0.66      0.74      0.70        77
           1       0.80      0.73      0.76       107
```

Figure 2: Predictions from transformed testing data

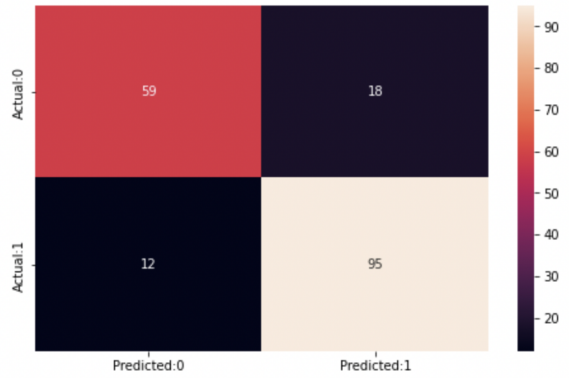

Figure 3: Confusion matrix for normal classifier

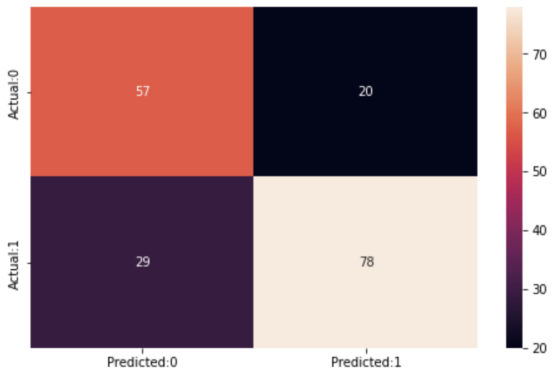

Figure 4: Confusion matrix for differential private classifier

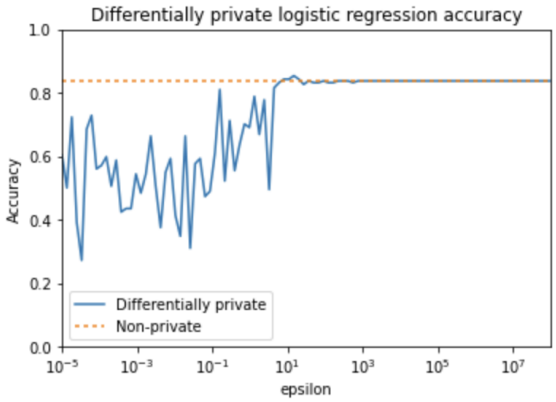

Figure 5: Accuracy comparison of Non DP and DP classifier

6. **Limitations of the model based on your findings from Part 1 of this assignment.**
The following factors might limit or degrade the performance of the model:

- Including too many input variables might can dilute true associations and lead to large standard errors with wide and imprecise confidence intervals.
- If input variables are highly correlated with one another, then the effect of each on the regression model will become less precise.

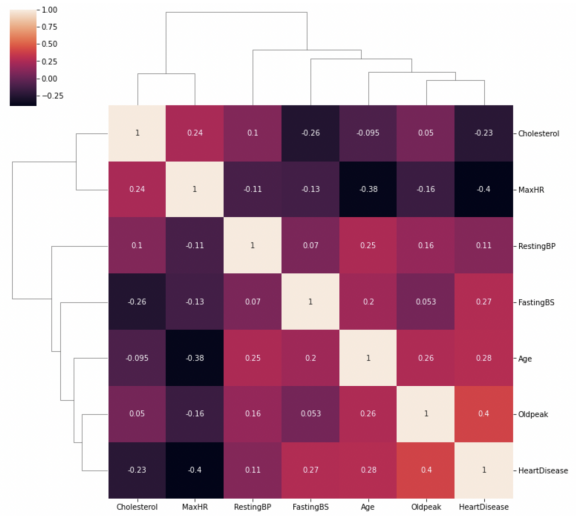

Figure 6: Heatmap for understanding the correlation between different features of the dataset

- For continuous data (e.g., age), dividing the subjects into categories (e.g., age >50 years vs. age <=50 years) is not a good practice as a part of the information might be lost.
- Regression equation derived from this specific set of patients might not apply to patients with different characteristics.
- Making the classifier differentially private trades-off with the performance metric of the classifier, as can be seen in figure 5