ECSE 557

Intro to Ethics of Intelligent Systems

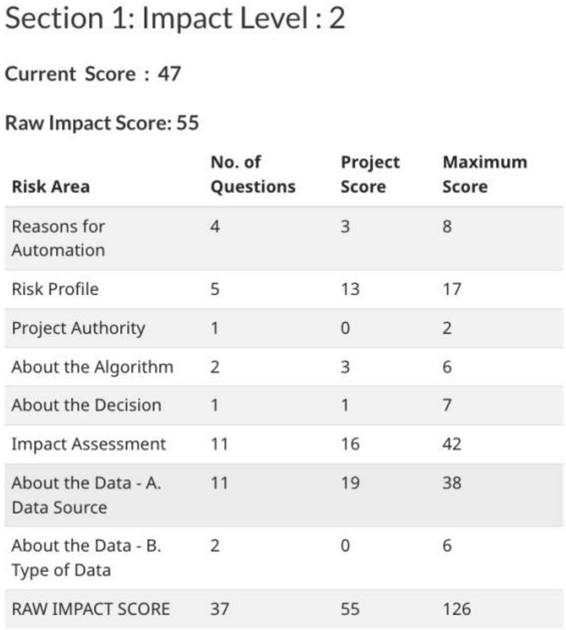
Assignment 4:

Putting it all together

CHIKHAOUI Hamza 260912960

Winter 2024

1. See the “AIA-en.pdf” file submitted alongside this report.
2. After conducting the AIA, we see that we have achieved an impact level of 2, with an overall current score of 47. A more precise breakdown of the raw impact and the mitigation score of our model can be found below.



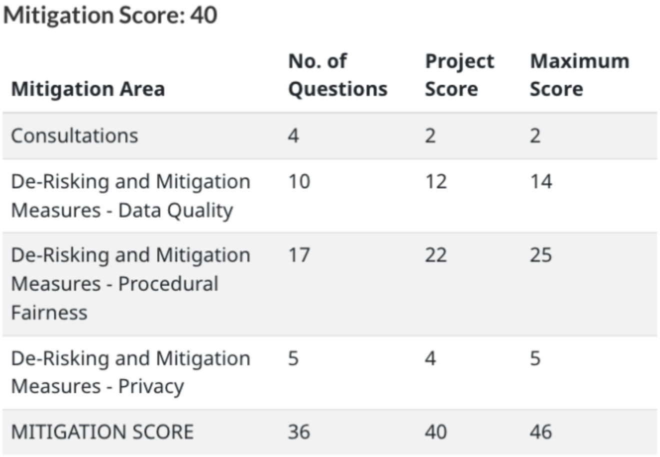
Section 1: Impact Level : 2

Current Score : 47

Raw Impact Score: 55

Risk Area	No. of Questions	Project Score	Maximum Score
Reasons for Automation	4	3	8
Risk Profile	5	13	17
Project Authority	1	0	2
About the Algorithm	2	3	6
About the Decision	1	1	7
Impact Assessment	11	16	42
About the Data - A. Data Source	11	19	38
About the Data - B. Type of Data	2	0	6
RAW IMPACT SCORE	37	55	126

Figure 1 : Raw Impact Score Summary



Mitigation Score: 40

Mitigation Area	No. of Questions	Project Score	Maximum Score
Consultations	4	2	2
De-Risking and Mitigation Measures - Data Quality	10	12	14
De-Risking and Mitigation Measures - Procedural Fairness	17	22	25
De-Risking and Mitigation Measures - Privacy	5	4	5
MITIGATION SCORE	36	40	46

Figure 2 : Mitigation Score Summary

When it comes to the raw impact of our algorithm, we can see that we achieve a score of 55 out of a maximum of 126. This in turn classifies our model in the second level in terms of impact, meaning that our model has a moderate impact. This score is attenuated thanks to the relatively high mitigation score we have achieved.

When completing the AIA, multiple mitigation strategies were considered to reduce the risk level of TriageAssist. The achieved mitigation score is of 40 for a maximum score of 46. This in turn may be explained by the fact that this section of the questionnaire was answered based on what Prioritize Inc. should do as a company ideally. Therefore, this performance takes into account most of the possible mitigations measures that could have been taken into account.

a. Consultation:

We consulted internal stakeholders, particularly experts in client experience management and data governance, to ensure the secure storage of user data thereby safeguarding patient privacy, and to incorporate feedback from our clients such as hospital nurses, doctors, and administrators.

Additionally, we engaged external stakeholders, specifically hospital employees, members of the civil society and members of academia who utilize TriageAssist, to enhance our algorithm. We are also note that the generation of the AIA in no way implies to post the algorithm, data, or any part of the system to the open government platform. Therefore, our assessment does not constitute any breach of privacy to the patients' data.

b. Data Quality:

The respect of privacy, and making sure that all patients get the care that they deserve is at the center of Prioritize's mission.
This is why we have implemented multiple processes to test our datasets against biases and unexpected outcomes. We used aif360 to calculate fairness metrics to determine the extent of the disadvantage that women face compared to men when using the classifier, in order to reduce any existing bias. While the exact process is not publicly available, the tools used are. We have also implemented differentially private models from the diffprivlib library to ensure the privacy of the patients. However, apart from an in-depth data analysis, no additional processes have been implemented to resole any data quality issues on the existing base dataset. This was done in part because this section of our dataset is publicly available alongside the sources from which that dataset was built, thus allowing for reproductible results.
As we will be collecting new data from patients after the deployment of our algorithm, we intend the data collection process to drastically comply with stringent quality standards. We have also undertaken the Gender Based Analysis Plus to ensure there was no discrimination based on gender.
The machine learning engineers, data scientists and project managers are all held accountable for the design, development, maintenance and improvement of the system.
We also have incorporated a documented and publicly available process to manage the risk of outdated or unreliable data being used in making a classification decision. This would be done via a regular retraining of the classifier, and the setup of filters to detect unusual samples during data collection.

c. Procedural Fairness:

We have achieved a project score of 22 out of a maximum of 25 in the procedural fairness category. This is a result of our commitment to transparency, and the continuous improvement of our model.

This is also why we have implemented an audit trail that records all the recommendations and decisions made by the system. This audit trail also identifies the relevant authorities and delegated authorities competent as per the legislation in place. All key decision points are identifiable in the audit trail, and are within the automated system's logic linked to the legislation pertaining to the healthcare sector. While the audit trail shows the identity of the authorized decision maker, the algorithm as of now is not able to produce reasons for its decisions or recommendations as these would hardly be interpretable by users. However, Prioritize controls the access to the classifier, with hospitals operating the software at the discretion of their managing teams. The model also has a mechanism to capture the feedback by the users, and this feedback is communicated to our design teams to ensure the constant improvement of our product.

At any point during the triaging process of patients, health care practitioners are able to disregard the output of the algorithm as Triage Assist only provides a complementary recommendation. Therefore, it is possible at all times for humans to override system decisions. Those events are then placed in logs to be able to be analyzed layer.

d. Privacy:

We prioritize our patients' privacy and have implemented differential privacy techniques to ensure the security of their medical data. Our approach involves handling medical data with great caution, obtaining informed consent from patients before any utilization or storage of their data. Furthermore, as a closed system, we do not utilize APIs, and TriageAssist operates without requiring an internet connection. We have also generate k-anonymous and l-diverse versions of our dataset to protect the privacy of the users,

To put it in a nutshell, we have employed multiple mitigation techniques, such as audit trail, periodic retraining, data filters, AIF360 fairness metrics, differential private libraries, k anonymity and l diversity, Gender Based Analysis plus, and a rigorous and publicly available documentation to ensure high standards in terms of consultations with the clients, data quality, procedural fairness, and privacy.

This in turn explains our high performance in all of the mitigation areas.

3. See pdf

4. We have decided to get rid of the procedural fairness mitigation technique to study its impact on the overall score of our model. We achieved the following results

Table 1 : AIA Assessment with Procedural Fairness

Impact Level	2
Current Score	47
Raw Impact	55
Mitigation Score	40

Table 2 : AIA Assessment without Procedural Fairness

Impact Level	2
Current Score	55
Raw Impact	55
Mitigation Score	18

The exact method described to compute the new current score is available in the references section of this report.

We see that while our model is still under the scope of a level 2 project, the score percentage range of our project went from 31.30% to 43.6%, which gets us substantively closer to a level 3 project. Our mitigation score has dropped from 40 to 18, while our current score increased from 47 to 55. Reflecting on these key metrics, one can comprehend the significant impact the removal of only one mitigation strategy can have on the overall impact of an AI system.

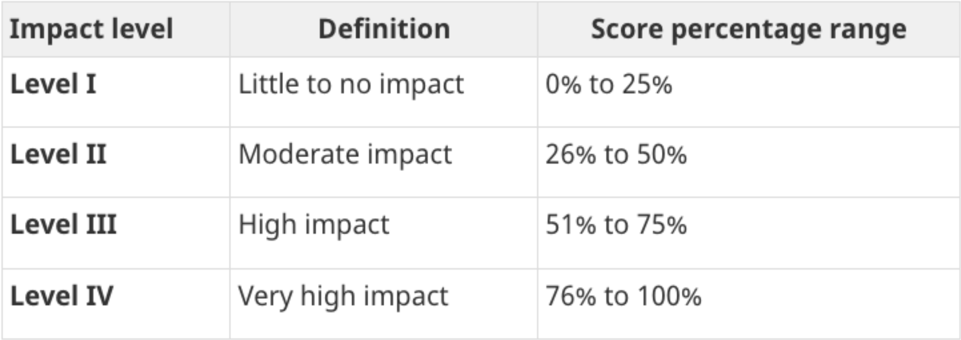
One can also nuance the previous statement by arguing that not all mitigation techniques have the same weight. Indeed, the consultation mitigation methods do not carry as much weight as the procedural fairness ones, as their maximum achievable weight is of 2, whereas the maximum achievable weight for procedural fairness is of 25. This does not however mean that consultations or privacy mitigation techniques should be disregarded. On the contrary, both of these mitigation techniques can have direct effects that can ripple deeply on the company’s reputation, or on the health and integrity of the patients / health practitioners.

References:

The current score is determined as follows:

- **If the mitigation score is less than 80% of the maximum attainable mitigation score (here 36.8),** the current score is equal to the raw impact score
  or
- If the mitigation score is 80% or more than the maximum attainable mitigation score, 15% is deducted from the raw impact score to yield the current score

Table 3 : Impact level calculation method



Impact level	Definition	Score percentage range
Level I	Little to no impact	0% to 25%
Level II	Moderate impact	26% to 50%
Level III	High impact	51% to 75%
Level IV	Very high impact	76% to 100%

This yields in our case 55/126 = 43.65%
As per the following :

[1] T. B. of C. Secretariat, "Algorithmic Impact Assessment Tool," *Canada.ca*, Apr. 25, 2023. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html