Group Project

2IX30 Responsible Data Science

CITI Course Deadline: February 20, 23:59 CET
Group Registration Deadline: February 27, 23:59 CET
Submission Deadline: April 24, 23:59 CET

Overview In this assignment you will build a prototype for a machine learning based decision-support tool at an Intensive Care Unit (ICU). You will make use of MIMIC-III, a real-world database on critical care.

Group Registration

You are responsible for registering in assignment groups of **at most 5 students** on Canvas. The group registration deadline is **February 27, 23:59 CET**.

Submission

The deadline for this assignment **April 24, 23:59 CET**. Please submit the following files on Canvas:

- *Report (pdf).*
- *Code (zip).* Source code of your experiments (e.g. juypter notebooks, or .py files).
- *Results (zip).* Any other raw results that are not included in your report.

Report Guidelines

- There is no page limit. As an indication, **15** pages is common (*excluding* front page, table of contents, references, and other appendices).
- Throughout the project you may move back and forth between the different stages of the development process. However, the report should **not** be a chronological report of your activities. Aim to structure your report as described in the assignment. You are welcome but are not required to add an appendix with your reflections about the development process and lessons learnt.
- Make sure you understand what you wrote. Spell check, grammar check, and proof read the document before handing it in.
- Each figure and table should be numbered and accompanied by a caption text that explains what the reader sees. Refer to figures and tables in the text by using their numbers, e.g., *"Figure 1* shows...". A figure caption is centered *under* the figure; a table caption is centered *above* the table.
- Clearly reference any resources you have used in an appropriate manner using proper citations (used software/libraries, papers, collaboration with other groups if any).

  Do not copy whole sentences from websites, articles, books, or your peers. Reports will be checked for plagiarism. Procedure requires that plagiarism is reported to the examination committee.

Grading

The total number of points that can be earned with this assignment is 100. Your grade is equal to the total number of points you have earned, divided by 10.

- 90 points can be earned by completing the tasks.
- 10 points are reserved for the overall presentation quality of your report. In case the work breakdown and/or individual reflections are not included in the report, 5 points will be subtracted from the points you have earned for the quality of your report.

Each member of a group will, in principle, get the same grade on the assignment. In case some group members contributed much more or much less than others, this may be reflected in the grade accordingly. You should provide this clarification in the individual reflections appendix of your group report. A more detailed grading rubric will be made available on Canvas.

Please note that it is expected that throughout the project, you will move back and forth between the different steps of the development process. Take this into account when dividing tasks.

Tools and Resources

You are encouraged to use GitHub (or similar) for ease of version control and project management. See this blog for an introduction to Github.

There are several Python libraries that may be useful for this assignment, including (but not limited to):

- *Data pre-processing.* numpy, pandas, scikit-learn
- *Visualization*: matplotlib, seaborn
- *Machine learning.* scikit-learn, xgboost
- *Fairness.* fairlearn, AIF360
- *Explainability.* interpretml, shap, AIX360

When choosing a library, please take into consideration that the available documentation varies considerably between libraries. In particular, fairness and interpretability libraries are substantially less mature than more established libraries such as `numpy` and `scikit-learn.`

Assignment Description

Scenario

The goal of this assignment is to develop a prototype for a *mortality prediction* model that is to be used as a decision-support tool for critical care physicians at Beth Israel Deaconess Medical Center.

Currently, the physicians at the hospital rely on the sequential organ failure assessment (SOFA) score for identifying patients which suspected infection who are at great risk of mortality. SOFA employs six criteria reflecting the function of an organ system (respiratory, cardiovascular, renal, neurological, hepatic and haematological) and allocates a score of 0–4 to each, depending on the value of relevant lab measurements. The SOFA score is a simple tool with a high sensitivity (true positive rate). However, the predictive performance of the tool is lacking. Moreover, the ICU director is worried that diverse patient populations are not accurately represented, as the tool does not take into account patient demographics nor differing infection sources.

In the past few years, the hospital has collected a vast amount of electronic health records (EHR). The ICU director wonders if this data can be used to make better mortality predictions. Your team is asked to explore the utility of machine learning for this task. The model is to be used as a decision support tool for physicians to determine appropriate levels of care and discuss expected care outcomes with patients and their families.

Dataset

MIMIC-III is a freely available database developed by the MIT Lab for Computational Physiology, comprising of de-identified health data associated with approximately 60,000 intensive care unit admissions. The database includes demographics, vital signs, laboratory tests, medications, and more.

As the data set contains clinical data of **real** people, appropriate care must be taking when handling the data set. Instructions on how to request access to the MIMIC-III database and the pre-processed data set are available on Canvas.

Please ensure that the data set is accessible to no one but yourself, e.g., by storing the data locally on your laptop. In particular, **make sure that any shared repositories (GitHub, Dropbox, etc.) within your group do *not* contain MIMIC-III data.** Once you have submitted your final report, the data must be removed from your laptop.

For this assignment, you can use a (partially) pre-processed data set extracted from the MIMIC-III database. The code to further pre-process this dataset is available on Canvas.

SOFA scores

In addition to a basic pre-processing steps, you are provided with a function to compute the SOFA scores for each patient in the data set. This will allow you to compare your prototype with SOFA.

Suggested Reading

- S. Wang, M. B. A. McDermott, G. Chauhan, M. C. Hughes, T. Naumann, and M. Ghassemi. *Mimic-extract: A data extraction, preprocessing, and representation pipeline for MIMIC-III.* CoRR, abs/1907.08322, 2019. Available at: `https://arxiv.org/abs/1907.08322`

- S. R. Pfohl, A. Foryciarz, and N. H. Shah. *An empirical characterization of fair machine learning for clinical risk prediction.* Journal of Biomedical Informatics, 113:103621, Jan. 2021. Available at: `https://www.sciencedirect.com/science/article/pii/S1532046420302495`

- Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. *Dissecting racial bias in an algorithm used to manage the health of populations.* Science, 366(6464):447–453, Oct. 2019. Available at: `https://www.science.org/doi/10.1126/science.aax2342`

- Rudin, C. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence, 206–215, May 2019. Available at: `https://doi.org/10.1038/s42256-019-0048-x`

1 Problem Understanding (20 points)

- *Problem Background.* Describe the problem background and the objective of the envisioned system (purpose, intended use case, expected benefits).
- *Risk Assessment.* In your risk assessment, include the following:
  - All stakeholders of this scenario, including a short description of each stakeholder.
  - The potential benefits, including the value it brings and to which stakeholder it is beneficial.
  - The risks of the development and usage of the system. In your report, include a table in which you explain for each risk: (1) the potential harm, (2) which stakeholder is impacted, (3) which (moral) value is at stake, (4) the severity of the harm, and (5) the likelihood of the harm.
  - A summary of the technical mitigation strategies that you use to mitigate the risks.
- *Main requirements and real-world success criteria.* Based on the problem background and risk assessment, define a list of main requirements and success criteria of the project.
- *Machine Learning Task Formulation and Technical Success Metrics.* Describe how the business problem was translated into a machine learning task formulation, including the data that is used as input and the target variable.

  Note that success criteria can include both quantitative metrics (e.g., predictive performance metrics, fairness metrics, training time) and more qualitative criteria (e.g., interpretability).

  Additionally, select relevant qualitative or quantitative success metrics:

  - Predictive performance metric(s) (e.g., accuracy, area under the ROC curve, false positive rate, false negative rate);
  - Fairness metric(s) (e.g., demographic parity, equalized odds);
  - Interpretability criteria (e.g., local/global, post-hoc/directly interpretable);
  - Privacy criteria and/or metric(s) (e.g.,)
  - Accountability criteria (e.g., the required documentation)
  - Other performance metrics (e.g., training time)

  For each of the metrics/criteria, motivate why they are relevant for the problem at hand.
- *Assumptions.* If this prototype was developed for a real-world scenario, we would advise you to incorporate perspectives of a diverse set of stakeholders throughout the development process. As this is not possible within the context of this course, we instead ask you to identify assumptions you have made throughout the problem understanding and how you would validate these assumptions in practice.

  Include your list of assumptions and suggested validation approaches as an **appendix** in your report.

2 Data Understanding & Preparation (20 points)

- *Data description.* In your report, briefly describe the original data, including the number of instances, the features, and how it was collected and processed.
- *Data exploration.* Explore the data and report any interesting findings, such as missing values, data distribution, pairwise correlations, etc.

- *Data pre-processing.* Pre-process your data set such that it is suitable for the machine learning algorithms you are considering. In your report, briefly describe the steps of your (final) pre-processing pipeline and explain your decisions. Additionally, briefly describe the final pre-processed data, including the number of instances and the included features.
- *Data Sheet.* Prepare a data sheet of the pre-processed data (see Gebru et al. [2018]) and include it as an **appendix** in your report.

3 Modelling (20 points)

- *Candidate algorithms.* Describe the machine learning algorithms you will try (e.g., logistic regression, decision tree, random forest, XGBoost). For each algorithm, describe possible advantages and disadvantages, given the problem requirements.
- *Model Selection.* Use the algorithm(s) to train machine learning models (one for each algorithm). Describe your model selection pipeline, including the model selection procedure (hyperparameter tuning procedure, cross-validation, etc.) and the evaluation metric(s) that you use during model selection.
- *Choose a decision threshold.* Based on the criteria you have defined in Task 1 (Problem Understanding), choose an appropriate decision threshold for...

  - each of the trained models
  - the SOFA scores

  *Hint: use the **training** set to choose a threshold and the **validation** set evaluate each of the thresholds (so **not** the test set).*
- *Choose a final model.* Use the criteria you have chosen in Task 1 (Problem Understanding) to evaluate your trained models with the selected thresholds and choose one final model.

  Hint: use the validation set to perform the final model selection.
- *Results.* Present and interpret the results from the model selection.

4 Evaluation (20 points)

- *Quantitative evaluation.* Thoroughly evaluate the final model on the **test** set, given the metrics you have selected in Task 1. In particular, compare your final model with the SOFA tool.

  In your report, describe your evaluation approach and present and interpret the results.
- *Qualitative evaluation.* Use intrinsic interpretability or post-hoc explanations of the chosen model to explore and evaluate its behavior. Which features are primarily used to make predictions? Does that make sense, e.g., compared to the features used by SOFA? Present and interpret the results.
- *Model Card.* Prepare a *Model Card* of your final model (see Mitchell et al. [2019]) and include it the **appendix** of your report.

5 Conclusion and Discussion (10 points)

- *Conclusions.* Briefly recap what you have done in this project, highlighting important accomplishments or results. Describe what your results mean given the original problem statement, i.e., to what extent you have “solved” the problem.

- *Limitations.* **Critically** discuss the limitations of your project. In particular, reflect on ethical implications. Was the data appropriate for the purpose? Is the implemented prototype robust/accurate/fair/stable/useful/etc.?
- *Future work.* What still remains to be done? What do you think are the next steps? In particular, describe how you would further evaluate your model before it is put into production.

6 Individual Reflection & Work Breakdown

- *Individual Experience Documentation.* Each group member must write a paragraph (approximately 100 - 200 words) in which they reflect on their individual experience of the project. Some questions to consider: what did you learn from the project? Were there any pitfalls? What were the strengths or weaknesses of how you approached the project? What would you do differently next time? Were there any parts that were particularly interesting (or frustrating)?

  Include the individual experience documentation as an **appendix** in your report.
- *Work Breakdown.* Prepare a work breakdown that indicates who contributed to which parts of the project and (approximately) how many hours were invested for each task.

  Include the work breakdown as an **appendix** in your report.

References

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford. Datasheets for Datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018. URL `http://arxiv.org/abs/1803.09010`.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 220–229, oct 2019. doi: 10.1145/3287560.3287596. URL `http://dx.doi.org/10.1145/3287560.3287596`.