Group Project

2IX30 Responsible Data Science

Submission Deadline: 24 March 2025, 23:59 CET

Overview In this assignment you will build a prototype for a machine learning based decision-support tool for Intensive Care Units (ICUs).

Submission

The deadline for this assignment **24 March 2025, 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 minimum or maximum number of words nor a page limit. As an indication, **10** pages should be sufficient (*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. The report should **not** be a chronological report of your activities. Aim to structure your report as described in the assignment.
- Make sure you understand what you wrote. Spell check, grammar check, and proofread 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.

Plagiarism Policy

You must reference any resources you have used in an appropriate manner using proper citations (used software/libraries, papers, online resources, collaboration with other groups if any).

- Feel free to use any resources you can find on the internet (including the WiDS 2020 Kaggle page), but **do not copy whole sentences from websites, articles, books, or your peers** without proper citation. Reports will be checked for plagiarism. We are required to report plagiarism to the examination committee.
- **Be transparent concerning the use of large language models (LLM), like ChatGPT**. You are allowed to use LLMs for editing or polishing text you have written yourself. *You must indicate the use of an LLM in your work breakdown, including an example prompt.* **Do not include texts entirely generated by an LLM.** We will investigate potential violations and treat these with gravity similar to reports flagged for potential plagiarism.

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 would 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, polars
- *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. The challenge is to create a model that uses data from the first 24 hours of intensive care to predict patient survival.

Currently, physicians often rely on APACHE ("Acute Physiology and Chronic Health Evaluation") for identifying patients at risk for mortality. APACHE is a scoring system assessing severity of illness and prognoses of ICU patients. The scoring system has been improved over time, with APACHE II being released in 1985, APACHE III in 1991, and finally APACHE IV in 2006. The APACHE IV score is computed based on a patient's age, several physiology measurements, and chronic illness indicators. While APACHE IV has been shown to perform well on various predictive performance metrics, the underlying model assumes linear relationships between the variables and mortality. Moreover, the scores have not been validated across demographics and hospitals in terms of predictive performance and calibration.

An international research project has been established to explore how mortality prediction can be further improved. In the past few years, several hospitals have collected a vast amount of electronic health records (EHR). Your team is asked to explore the utility of machine learning for mortality prediction using this data. 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

The data set that will be used in this assignment is collected from MIT's GOSSIS community initiative and includes more than 130,000 hospital Intensive Care Unit (ICU) visits from patients, spanning a one-year time frame.

You can retrieve the data set through these steps:

1. Create a user account on `https://physionet.org`
2. Read and sign the data use agreement: `https://physionet.org/sign-dua/widsdatathon2020/1.0.0/`
3. Download the files at the bottom of this page: `https://physionet.org/content/widsdatathon2020/1.0.0/`

As the data set contains clinical data of **real** people, appropriate care must be taking when handling the data set. 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 the data set.**

To help you get started, we have provided a Jupyter notebook on Canvas which you can use to (partially) pre-process the data set for a machine learning task.

Suggested Reading

- Zimmerman, J. E., Kramer, A. A., McNair, D. S., and Malila, F. M. (2006). *Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients.* Critical care medicine, 34(5), 1297-1310. Available at: `http://www.jvsmedicscorner.com/ICU-Miscellaneous_files/APACHE%20IV.pdf`
- 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 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 the 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, the target variable, and technical success criteria. For each of the metrics/criteria, motivate why they are relevant to the problem.

  Note: technical success criteria can include both quantitative metrics (e.g., predictive performance metrics, inference time) and more qualitative criteria (e.g., interpretability).
- *Assumptions.* If this prototype was developed for a real-world scenario, we would advise you to incorporate the 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 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. *Note: it may not be possible to answer all questions in the dataset. Fill out the data sheet as best as you can with the information that is available to you.*

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 you have identified in Task 1 (Problem Understanding).

- *Model Selection.* Use the algorithm(s) to train machine learning models and, based on the criteria identified in Task 1, select a model.

  In your report, describe and motivate 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.

- *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 selected model with APACHE IV scores.

  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 model behavior. Which features are primarily used to make predictions? Does that make sense, e.g., compared to the features used by APACHE? 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, consider how you would further evaluate your model before putting it into production.

- *Recommendation.* Based on your analysis, provide a recommendation regarding the use of your model. Make sure to take into account all risks and benefits you have identified.

  - If you believe hospitals should use the system, argue why and describe the circumstances under which the system’s use would be appropriate.
  - If you believe hospitals should not use the system, argue why the risks outweigh the benefits or describe which conditions must be met before the system can be used.

6 Individual Reflection & Work Breakdown

- *Individual Self Assessment and Experience Documentation.* Each group member includes the following:

  - The **grade** they believe the project, as documented in the report, deserves. Include a brief motivation. *The rubric on Canvas can be helpful to perform this self assessment.*

  - A **personal reflection** 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.