Group Project Assignment Instructions

Machine Learning Applications for Health [COMP90089]

Mike Conway Brian Chapman Daniel Capurro

20th July 2024

Version: 1.4

1 Introduction

So far in this subject, you have been introduced to core concepts in health informatics and digital health, several different health data sources and their associated idiosyncrasies, standard procedures for preparing and cleaning health data with the goal of generating research-ready datasets, the development of digital phenotypes, and different machine learning methods and their application to health. This material all serves as background for the group project which makes up 50% of your total marks for this subject.

This is a group assignment and **groups will be randomly allocated in Canvas**. This group assignment forms the main summative assessment for the subject.

The assignment requires that you *design*, *implement*, and *evaluate* a machine learning project that addresses a *relevant health problem*.

The project has three parts:

1. A **project proposal** worth 10% of your marks for the subject
2. A **project report** worth 39% of your marks for the subject
3. A **team assessment** worth 1% of your marks for the subject

2 Project Proposal

The first deliverable, consists of a *project proposal* equivalent to approximately 500 words. This proposal will constitute 10% of your total marks for the subject. You will be required to:

- Briefly define the health problem your group chose and demonstrate its significance
- Describe the data processing and digital phenotyping approaches you intend to follow
- Describe the machine learning and evaluation strategies you will implement, including metrics of success

Regarding data, while your group is free to choose any health-related data source, given that MIMIC data is used extensively in the subject, it is expected that most groups will elect to utilize MIMIC as the primary data source for their assignment. However, if you decide to use a different dataset (i.e. not MIMIC) then that dataset should be publicly available.

For this project proposal, we ask you to answer the questions below, in a succinct way.

1. What is the clinical question/problem your group will tackle and why is it relevant? Please include information about the magnitude of the problem and make sure your claim is supported by references to the literature (up to 100 words).
2. Which data source(s) and information will you use? What is your plan for data processing and digital phenotyping? Remember for a digital phenotype it is important to list the criteria and describe which data elements you will use to find patients that meet those criteria (up to 200 words).
3. What is your machine learning approach and methodology? What are your metrics of success and expected outcomes? (up to 300 words)

Our expectation is that you will cite at least five references in writing your project proposal. Note that it does not matter which citation style you select, as long as it is consistent throughout the project proposal.

3 Final report

The second deliverable consists of a *project report* equivalent to approximately 2,000 words. This report will constitute 39% of your marks for this subject. Your report should:

1. Introduce the health problem your group has decided to tackle and its significance (building and expanding from your project proposal document);
2. Describe the methodology your group has followed, including:
   (a) Data processing and digital phenotyping approaches you have followed
   (b) Machine learning and evaluation strategies you have implemented
3. Describe and discuss the obtained results, their implications, how they align with the literature and potential future steps.

Each group should include, at the end of their report, a table of contributions made by each group member using the **CRedIT - Contributor Role Taxonomy**[1]

Note that in assessing your project, we am more interested in your critical analysis of methods and results than the raw performance of your models. It may be that you do not arrive at a definitive answer to your research question. However, you should analyse and discuss your (possibly negative) results in depth. You will be assessed on the quality of your report, **not** your code. However, your code should be sufficient to replicate the results provided in your report (i.e. in principle, it should be possible to run the code you supply to generate the results provided in your report). You are not required to implement algorithms from scratch. Using existing library implementations of algorithms is encouraged.

The report will consist of:

1. An introduction to the research question you have decided to address and its significance
2. A description of the resources and methodologies adopted, including evaluation strategies
3. A description and discussion of the results obtained, the implications of these results, how well your results align with the research literature, and future steps

You should write the report using the JAMIA (*Journal of the American Medical Informatics Association*) format. A description of the format can be found here, but briefly, reports should

[1] CRedIT Taxonomy: https://credit.niso.org

consist of:

- **Structured abstract** – Write a short, structured abstract of up to 250 words containing the headings *Objective*, *Materials and Methods*, *Results*, *Discussion*, and *Conclusion*)
- **Introduction** – Provide a short description of the problem and data set, and the research question addressed. The Introduction section should include a summary of some related literature
- **Methods** – Explain the primary machine learning methods that you have used in your project with appropriate references. This section should include a description of your dataset and an ethics statement [Note that the ethics statement does not count towards the word count for the assignment]
- **Results** – Present your results in terms of evaluation metric(s) and, ideally, illustrative examples. Use of tables and figures is highly recommended
- **Discussion & Conclusion** – Contextualize and reflect on your results in the context of your research question. What has been learnt? What are the limitations of your project?
- **References** – a specific referencing style is not required, but consistency in how you format references is important
- **Appendixes** – including all code used

The Methods section should include an ethics statement [use the heading **Ethics Statement**]. In most circumstances, ethics statements can be short and should typically discuss (a) whether the data is public; (b) any plans to redistribute data generated during the project; and (c) any sensitivity involved with the data.

Note that the Ethics Statement will not count towards the word count

Our expectation is that you will cite 15 to 30 references in writing your report. Note that it does not matter which citation style you select, as long as it is consistent throughout the report.

You can write the report using Microsoft Word, LaTeX, or Markdown. Templates are available for LaTeX[2] and Markdown[3]. Note that you can use as many tables and figures as you like, but try and keep the word count around 2,000. Further, please make sure that any figures used in the main body of the report are referenced in the text and discussed. Note that there is no word limit on material in the appendixes (please include all the code you use as an appendix).

Your report will be assessed against the following criteria:

- Clarity of expression and cohesion
- Adequate contextualisation of the health problem in the literature
- Adequate use of visual aids (plots, figures, flowcharts, tables)
- Adequate methodological steps (i.e. sufficient detail to – in principle – replicate the work)
- Adequate discussion of results, implications, and future directions

Finally, in previous years we have seen many projects that run dozens of models and simply report which one performs best, with little analysis or discussion of the problem being studied. It is almost certainly a better strategy to evaluate fewer models — for example, a more interpretable model, a "black box" model, and a baseline model— combined with a deep analysis of the particular health-related problem you have selected.

[2] LaTeX template: https://github.com/emir-munoz/amia-paper-template
[3] Markdown template: https://github.com/erictleung/jamia-markdown-template

4 Team Assessment

On Canvas you must complete the team assessment assignment for the group project. In the team assessment, list each member of your team. For each team member assign a score from **1** (*very poor team member*) **to 5** (*excellent team member*) rating their contribution to the project. Justify your score with a brief sentence for each team member.

5 Uploading to Canvas

Please submit both the project proposal and final project report files as PDFs with your group number, the names of your group members, the student ID numbers of your group members, and your report's title clearly indicated on the title page. Please use page numbers in your report.

Please title your file for the proposal as follows, replacing {GROUPNUMBER} with your group number:

`ML4Health_proposal_{GROUPNUMBER}.pdf`

Please title your file for the final project as follows, replacing {GROUPNUMBER} with your group number:

`ML4Health_report_{GROUPNUMBER}.pdf`

6 Example Projects from Previous Year

Previous project titles include:

- Using blood glucose data to improved predictions of mortality for sepsis patients in the intensive care unit (ICU)
- Machine learning implementations on the COPD cohort of MIMIC-IV v2.2: K-means clustering and mortality prediction
- Diagnosing ventilator-associated pneumonia: a machine learning approach with MIMIC-IV data
- Predicting in-hospital mortality for ICU patients diagnosed with myocardial infarction: a machine learning approach using data from the first 24 hours of admission
- Mortality predictive analysis on acute pancreatitis in intensive care unit patients - MIMIC-IV dataset
- Prediction of adverse drug events due to drug-drug interactions in the intensive care unit
- Risk prediction modeling for ventilator-associated pneumonia (VAP) using early indicators in intensive care unit (ICU) patients

7 Academic Misconduct

We will be checking submissions for originality and will invoke the University's Academic Misconduct policy where inappropriate levels of plagiarism are deemed to have taken place. Content produced by generative AI (including, but not limited to, ChatGPT) is not in a straightforward sense your own work, and submitting such content may be treated as a case of academic misconduct, in line with University of Melbourne policy.

8 Intended Learning Outcomes and Generic Skills

This project is associated with the following Intended Learning Outcomes:

- Describe the application of Artificial Intelligence concepts in the context of health problems
- Demonstrate familiarity with challenges associated with health data and data modelling
- Design, implement, and evaluate an AI system addressing a healthcare problem
- Critically assess Artificial Intelligence applications for health