GROUP PROJECT

This assessment provides you with an opportunity to reflect on concepts in machine learning in the context of an open-ended research problem, and to strengthen your skills in data analysis and problem solving. That is, the idea behind the project is for you to correctly implement general principles of statistical machine learning, while exploring data and algorithms of your interest. The goal of this project is not to obtain the best performance metric (e.g., accuracy) per se, but to perform different steps of machine learning in the proper way, according to what you have learnt in this subject. You should be clear about what is the research question in your project, what you plan to try, and what insights you might be planning to get. Then in terms of the results you get, you should discuss what worked or what did not work, and explain the possible reasons in light of what you learnt in class.

1. Overview

On Tuesday September 17, 11.59pm or later, you should submit through Canvas (not graded):

- Your project plan (see ProjectPlan.docx). The plan (or any subsequent change) will be reviewed, to ensure that all plans have a similar level of complexity. If your plan is not reviewed, it might likely not be at the level of what we expect, which will end up impacting your grade. Thus, it is in your best interest to have your plan reviewed.
- Your group agreement (see GroupAgreement.docx). While not a formal legal contract, completing the agreement together is a helpful way to open up communication within your team, and align each others' expectations.

On Friday October 11, 11.59pm, you should submit through Canvas:

- A written 4-page research report in PDF format (details in Section 2).
- A ZIP archive of your source code with a complete and correct implementation of your plan. This should also contain a Readme.txt file (describing in just a few lines what files are for, and how to run the code) and any scripts for automation. We are unlikely to run your code, but we may in order to verify the work is your own, or to settle rare group disputes.
- You may include Slack/Github logs if you wish. We require your team to meet at least 3 times and have included a template for recording high-level minutes if you wish, which can be included.
- Do not submit data!

2. Report

A 4-page research report should be submitted through Canvas. In the report, we will be interested in seeing evidence of your thought processes and reasoning for choosing one method over another, in a correct fashion. The report should include the following content:

1. Introduction: A brief description of the problem, dataset, notation and the research question addressed.
2. Literature review: a short summary of some related literature, including the data set reference and at least two additional relevant research papers of your choice (One of those papers is related to one of the algorithms you will implement. See question 6 in the project plan).
3. Methods: Description of the proposed methods (e.g., feature construction, preprocessing, 3 algorithms, cross-validation, hyperparameter tuning). Your description of the proposed methods should be clear and concise. You should write it at a level that a postgraduate student can read and understand without difficulty. For instance, for the research-paper algorithm, *please do not rewrite the complete description, but provide a summary* that shows your understanding and references to the relevant literature.
4. Results and Discussion: Experimental results (e.g., tables, charts) and a discussion addressing the differences in performance of different proposed methods. Provide your analysis and insights of why the proposed methods work/not work for the problem and dataset. Clearly discuss their advantages/disadvantages based on the understanding from the subject materials as well as in the context of your research question. Include also other alternative approaches you considered and why you chose your proposed methods over these. (While empirical evaluation should support your reasoning, your reasoning should go beyond just empirical evaluation — examples like "method A, got accuracy 0.6 and method B, got accuracy 0.7, hence I use method B", with no further explanation, will be marked down).
5. Conclusion: Clearly demonstrate your identified knowledge about the problem.
6. Bibliography as well as references to any other related work you used in your project. *We discourage using problems/datasets that you previously used or are currently using on other subjects, if you do so, you should clearly cite your own previous and current work and explain the differences in this project.* You are encouraged to use the APA citation style, but may use different styles as long as you are consistent throughout your report.

Report format rules. The report should be submitted as a PDF, and be no more than 4 pages, single column, A4 page size. The font size should be 10pt and margins should be between 1.5cm and 2cm on all sides. If a report is longer than 4 pages in length, we will only read and assess the report up to page 4 and ignore further pages. (Do not waste space on cover pages. References and appendices are included in the page limit — you do not get extra pages for these. Double-sided pages do not give you extra

pages — we mean equivalent of 4 single-sided. *Four pages means four pages total*. Learning how to concisely communicate ideas in short reports is an incredibly important communication skill for industry, government, and academia alike.)

3. Tentative Rubric

Completeness and correctness (Maximum = 10 marks)

10 marks	The implementations of the proposed methods (e.g., feature construction, preprocessing, 3 algorithms, cross-validation, hyperparameter tuning, experimental results, source code) are complete and correct.
8 marks	The implementations of the proposed methods have some minor issues in terms of completeness and/or correctness.
6 marks	Several aspects of the proposed methods' implementation are lacking and/or were incorrectly implemented.
4 marks	The implementations of the proposed methods have some serious issues in terms of completeness and/or correctness.
2 marks	The implementations of the proposed methods are incomplete and incorrect.

Clarity and Structure (Maximum = 5 marks)

5 marks	Very clear and accessible description of all that has been done, a postgraduate student can pick up the report and read with no difficulty.
4 marks	Clear description for the most part, with some minor deficiencies/loose ends.
3 marks	Generally clear description, but there are notable gaps and/or unclear sections.
2 marks	The report is unclear on the whole and the reader has to work hard to discern what has been done.
1 mark	The report completely lacks structure, omits all key references and is barely understandable.

Critical Analysis (Maximum = 10 marks)

10 marks	Proposed methods are well motivated and their advantages/disadvantages clearly discussed; thorough and insightful analysis of why the proposed methods work/not work for the problem and dataset; insightful discussion and analysis of alternative approaches and why they were not used.
8 marks	Proposed methods are reasonably motivated and their advantages/disadvantages somewhat discussed; good analysis of why the proposed methods work/not work for the problem and dataset; some discussion and analysis of alternative approaches and why they were not used.
6 marks	Proposed methods are somewhat motivated and their advantages/disadvantages are discussed; limited analysis of why the proposed methods work/not work for the problem and dataset; limited discussion and analysis of alternative approaches and why they were not used.
4 marks	Proposed methods are marginally motivated and their advantages/disadvantages are discussed; little analysis of why the proposed methods work/not work for the problem and dataset; little or no discussion and analysis of alternative approaches and why they were not used.
2 marks	Proposed methods are barely or not motivated and their advantages/disadvantages are not discussed; no analysis of why proposed methods work/not work for the problem and dataset; little or no discussion and analysis of alternative approaches and why they were not used.