1. Overview

This assessment provides you with an opportunity to reflect on concepts in machine learning in the context of an open-ended research project, 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 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. 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.

On May 27, 11.59pm AEST, you should submit through Canvas:

- A written 4-page research report in PDF format (*details in Section 3*).
- A ZIP archive of your *Python code or Jupyter notebook* and a Readme.txt file (describing in just a few lines 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.
- Do not submit data!

2. Defining your Methods

You will have to define different aspects of a machine learning project:

a. Which dataset(s) is(are) going to be used?
   - *Requirement:* Datasets should be already publicly available, since there is not enough time for you to collect data. Make sure accessing the data does not require special approvals that would take hours/days since this can delay your project. In your report, you will have to include the URL(s) of the dataset(s).
   - *Ideas:* Possible datasets can be found at Kaggle or the UCI Machine Learning Repository, but please read the requirements below.
   - *Requirement:* If you choose a tabular dataset, your dataset should have at least 5000 instances and 50 features. (For this, we consider the original features plus additionally constructed features. *See Section 2.c for more details.*) Of course, doing things such as duplicating instances or features, adding zeros or random data, in order to fit the required sizes is not allowed.
   - *Requirement:* If you choose an image dataset, your dataset should have at least 5000 images, and the original images should be of size at least 50x50 pixels. Of course, doing things such as duplicating images, scaling/resizing images, or adding extra pixels on the border, in order to fit the required sizes is not allowed.
   - *Advice:* Do not spend too much time on things such as "understanding the data", "computation or memory issues because the data is too big", etc. Only if you are already familiar with computer vision, brain data, natural language processing, big data, parallelism, etc. then you can make use of those things, but this will not imply that you will get a higher grade just based on that fact.
   - *Advice:* Try to use easy-to-understand datasets, and if the dataset is too large, you can use a small subset of the data since you will need to perform cross-validation and hyper-parameter tuning. (*See Sections 2.e and 2.f for more details*).

b. What is the problem and why this problem or dataset(s) is(are) particularly interesting for you?
   - For instance, your problem could be prediction of stock prices, or prediction of movie ratings, etc. Maybe you are interested in a problem because of your previous studies, your future career, a hobby, etc.

c. Which feature construction and preprocessing will be applied to the dataset(s)?
   - Simple methods (e.g., one-hot encoding, normalization/imputation, duplicate removal, sampling/filtering, merging/joining tables, image scaling/resizing, text processing) are acceptable for submission, but the more complex the methods, the higher your grade/marks will be.
   - *Requirement:* If your original dataset is a single data table, just reading the table and putting it through the machine learning algorithms (without any feature construction or preprocessing) is not allowed.
   - *Requirement:* If you choose an image dataset, just reading images and putting them through the machine learning algorithms (without any feature construction or preprocessing) is not allowed.
   - *Ideas:* The input data to the machine learning algorithms does not need to have the same format as the files in your dataset. You can perform complex manipulations such as:
     - If you have a file with sales including item, date, quantity and price, you can create a table where an item is a sample, possible features could be: average price 3 weeks ago, quantity sold 3 weeks ago, average price 2 weeks ago, quantity sold 2 weeks ago, etc.
     - If you have one file of patients, one file of lab visits, and another file of medical procedures performed on different dates and medical specialities, you can create one sample per patient, possible features could be: procedures per each medical speciality (one feature for each medical speciality), another feature could be lab visits, etc.
     - If you have images, you can consider rotations, adding noise, blurring images, adding occlusions, etc.
     - If you have 3 files: users, movies and ratings, you can create an incomplete matrix where user is a row, movie is a column, and each entry is the rating of a user for a movie.

d. Which 3 machine learning algorithms are going to be compared?
   - For instance, if you are performing classification, you should choose 3 classifiers, e.g., Naïve Bayes, Support Vector Machines and Decision Trees.
   - *Requirement:* Choose algorithms that cover some spectrum from high-bias/low-variance to low-bias/high-variance.
   - *Requirement:* Do not use the same algorithm with just different hyperparameters (e.g., SVM with C=1 and SVM with C=10). The 3 algorithms need to be different.

e. Which cross-validation technique is going to be used?
   - For instance, k-fold cross-validation, or several repetitions of train/test splits.
   - *Requirement:* Using a single test dataset is not allowed.
   - *Requirement:* If you choose train/test splits, make sure you have at least 5 different repetitions of the procedure, choosing different train/test datasets. Do not make much more than 5 repetitions since you will also need to do hyper-parameter tuning for which you will have to do nested cross-validation (cross-validation inside cross-validation).
   - *Requirement:* If you choose k-fold cross-validation, please use $k \geq 5$, but do not make k too large since you will also need to do hyper-parameter tuning for which you will have to do nested cross-validation (cross-validation inside cross-validation).
   - *Requirement:* Each of the 3 machine learning algorithms should follow the same cross-validation approach, in order to fairly compare the algorithms in your report.
   - *Ideas:* If you choose a temporal dataset, make sure you predict the future based on the past. You could divide the original dataset in 6 sequences/parts (one after the other). Then part 1 and 2 is one repetition. Part 2 and 3 is another repetition, ..., Part 5 and 6 is another repetition. Thus, you have 5 repetitions. Another way would be to sample randomly from the first half of the data to be the training data, and sample randomly from the second half of the data to be the test data. Then you would repeat this 5 times.

f. Which hyperparameter(s) is(are) going to be tuned for each of the 3 algorithms above, and what method is going to be used for the nested cross-validation.
   - *Requirement:* Every algorithm should have at least one hyperparameter to be tuned.
   - *Advice:* We recommend not to choose more than 3 hyperparameters per algorithm, otherwise it will become computationally demanding to perform hyper-parameter tuning.
   - *Advice:* Each hyperparameter should affect the results noticeably. For instance, if you try different values for a hyperparameter but classification accuracy does not change much, then we recommend choosing another hyperparameter.
   - *Requirement:* Choose which values you will consider for each hyperparameter. You should use at least 3 values and make sure that the value chosen most of the times by the tuning procedure is the middle one. For instance, if you choose to use C=0.1, C=1 and C=10 for Support Vector Machines, make sure that most of the times you choose C=1 when performing hyper-parameter tuning. Otherwise for instance, if C=0.1 was chosen most of the times, you could have considered lower values such as C=0.01. Similarly, if C=10 was chosen most of the times, you could have considered higher values such as C=100.

   - For nested cross-validation you can use for instance, k-fold cross-validation, or several repetitions of train/validation splits.
   - *Requirement:* Using a single validation dataset is not allowed.
   - *Requirement:* If you choose train/validation splits, make sure you have at least 3 different repetitions of the procedure, choosing different train/validation datasets.
   - *Requirement:* If you choose k-fold cross-validation, please use $k \geq 3$, but do not make k too large.
   - *Ideas:* If you choose a temporal dataset, you can follow a similar idea as in Section 2.e.

g. Which experimental results are you going to report?
   - For instance, learning curves, confusion matrix, precision, recall, accuracy, F1-score, root mean squared error, mean absolute error, etc.
   - *Requirement:* You should present at least 3 results, but the more results, the higher your grade/marks will be.
   - *Requirement:* "Loss with respect to number of iterations/epochs" is not acceptable. Note that learning curves (as described in our lectures) are different than "loss with respect to number of iterations/epochs".

3. Writing the 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. Student ID and Name.
2. Introduction: A brief description of the problem, dataset(s) and URL(s) and why this is particularly interesting to you. *For details, see Sections 2.a and 2.b.*

3. Literature review: a short summary of some related literature, including the dataset(s) reference(s), and at least one additional relevant research paper of your choice.
4. Methods: Description of the proposed methods (e.g., feature construction, preprocessing, 3 algorithms, cross-validation, hyperparameter tuning). *For details, see Sections from 2.c to 2.f.* 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.
5. Results and Discussion: Experimental results (e.g., tables, charts) and a discussion addressing the differences in performance of different proposed methods. *For details, see Section 2.g.* 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. 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).
6. Conclusion: Clearly demonstrate your identified knowledge about the problem.
7. 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.)

4. Tentative Rubric

Completeness and correctness (Maximum = 12 marks)

12 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.
9.6 marks	The implementations of the proposed methods have some minor issues in terms of completeness and/or correctness.
7.2 marks	Several aspects of the proposed methods' implementation are lacking and/or were incorrectly implemented.
4.8 marks	The implementations of the proposed methods have some serious issues in terms of completeness and/or correctness.
2.4 marks	The implementations of the proposed methods are incomplete and incorrect.

Clarity and Structure (Maximum = 6 marks)

6 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.
4.8 marks	Clear description for the most part, with some minor deficiencies/loose ends.
3.6 marks	Generally clear description, but there are notable gaps and/or unclear sections.
2.4 marks	The report is unclear on the whole and the reader has to work hard to discern what has been done.
1.2 marks	The report completely lacks structure, omits all key references and is barely understandable.

Critical Analysis (Maximum = 12 marks)

12 marks	Proposed methods are well motivated; methods' advantages/disadvantages are 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.
9.6 marks	Proposed methods are reasonably motivated; methods' advantages/disadvantages are 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.
7.2 marks	Proposed methods are somewhat motivated; methods' 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.
4.8 marks	Proposed methods are marginally motivated; methods' 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.
2.4 marks	Proposed methods are barely or not motivated; methods' 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.