School of Computing and Information Systems
The University of Melbourne
COMP90049, Introduction to Machine Learning, Semester 2 2025

Assignment 2: Group project

Released: Monday, September 15th 2025.

Due: **Group contract**: Monday, September 22nd 5pm
Project report and code: Friday, October 17th 5pm
Group reflection: Friday, October 24th 5pm

Marks: The Project will be marked out of 30, and will contribute 30% of your total mark.

1 Overview

This assessment provides you with an opportunity to apply and 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 correctly perform different steps of machine learning, according to what you have learnt in this subject. Based on the results you get, you should describe what worked or what did not work, and explain the possible reasons in light of what you learnt in class.

Each team is free to choose their own data sets and research questions. Section 3 provides detailed instructions and suggestions.

2 Assignment Structure and Deliverables

Here we describe all deliverables. Section 3 describes the content of the assignment.

2.1 Group Contract

Marks: 0 marks, -3 marks penalty if not submitted

Template: https://canvas.lms.unimelb.edu.au/courses/213823/files/folder/Assignments/Assignment%202?preview=24927493

Due: Monday, September 22th, 5 PM

Submit: single PDF file via Canvas (Assignment 2: Group Contract). One submission per group.

The group contract will lay out your team's goals, expectations, and policies for working collaboratively on the project (such as each team members tasks and contributions). It will be consulted when assessing the relative contribution of each group member to resolve any dispute.

A group contract template is provided. You are welcome to work with the provided template or customize it according to your preference. However, the overall content of the submitted contract should be in line with the template, and none of important details (such as responsibilities of each member) can be omitted. You may update your group contract throughout the semester (and upload a new contract), but proposed changes must be agreed to by all members. The contract won't be marked, but [-3] marks will be deducted from the final project mark if no meaningful contract is submitted by the deadline.

2.2 Report and Code

2.2.1 Report

Marks: 25

Template and length:ACL template (see below)

- **6-8 pages for groups of 3**
- **8-10 pages for groups of 4**

Due: Friday, October 17th, 5PM

Submit: single PDF file via Canvas/Turnitin (Assignment 2: Group Report). One submission per group.

Your report should consist of six to eight pages for groups of three people, and eight to ten pages for groups of four people. The report must use the ACL template linked below, which specifies font sizes, caption formatting, margins etc. Do not modify these settings (for example, to fit your content within the page limits). You can use a Word template (which describes the requirements in detail), or a Latex template. The templates are available here:

Word: https://canvas.lms.unimelb.edu.au/courses/213823/files/folder/Assignments/Assignment%202?preview=24928128

Latex: https://github.com/acl-org/acl-style-files/tree/master

Overleaf: https://www.overleaf.com/latex/templates/association-for-computational-linguistics-jvxskxpnznfj

The page limit includes all the text, captions, tables and images. References are **not** counted towards the page limit. While the template has Appendix section, **no appendices** are allowed. Tables and image content should be readable and sensible in size. All images and tables must have a meaningful, self-contained caption that captures the item's content, and must be referred to in the text. Reports must include at least one table and at least one image.

Marks will be deducted for non-compliance with these requirements, including violation of the length limit.

The group ID (as in Canvas) and all group members' names must appear on the first page after the title of the report.

The report must be submitted in PDF format. Docx or tex files are not accepted. If a report is submitted in any format other than PDF, we reserve the right to return the report with a mark of 0.

2.2.2 Code

Marks: 0

Template: None

Due: Friday, October 17th, 5PM

Submit: a single ZIP file via Canvas (Assignment 2: Code and Comments). One submission per group.

Python code or Jupyter notebooks necessary to reproduce the results in your report (model implementation, data processing, visualisation, and evaluation). If your code is **not** a notebook, please include a README file (describing in just a few lines how to run the code) and any scripts for automation. Jupyter notebooks should contain sufficient text comments to convey its general working and logic. Code will not be marked, but we may run the submitted code in order to verify the results in your report.

Do **not** submit your data but provide a link to its source in the report and README/notebook.

2.3 Group Reflection

Marks: 5

Template and length: 200-300 words, refer to the template below and in Feedback Fruits

Due: Friday, October 24th, 5PM

Submit: *Feedback Fruits available on Canvas (Assignment 2: Teamwork Evaluation).* Should be submitted individually by each group member.

Every team member needs to evaluate (1) their own contributions to the assignment and (2) the contributions of their teammates. This evaluation should align with the expectations you set in your submitted group contract.

Self-reflection

For self-reflection, submit a comprehensive reflective analysis (200-300 words) covering:

- Process Reflection: What worked well in your approach? What would you do differently?
- Learning Outcomes: Key insights gained about data science methodology?
- Team Dynamics: How did group collaboration contribute to or hinder the project?
- Technical Challenges: Major obstacles encountered and resolution strategies.
- Future Directions: How could this research be extended or improved?

Group evaluation

Provide a peer review of your collaborators. For each of the members of the group provide the following ratings (where 1 is ”terrible” and 5 is ”excellent”):

- Please rate their **general teamwork contribution** on a scale of 1 to 5.
- Please assess their contribution to the **coding part** of the project, considering their skills and expertise, using a scale of 1 to 5.

- Please assess their contribution to the **report writing** part of the project, considering their skills and expertise, using a scale of 1 to 5.
- Please assess their contribution in **sharing information and effectively communicating** with other team members, using a scale of 1 to 5.
- Please assess their **ability to prioritise tasks and meet the team milestones and deadlines**, using a scale of 1 to 5.

If any individual is identified as a non-contributor by their team members, the total marks for the report (25 marks) for that person may be adjusted accordingly. The Group Reflection mark (5 marks) is also calculated using the weight of marks received by peers.

3 Project development and requirements

Below we describe the steps which we expect you to take while developing your project, together with some ideas and requirements. Failure to follow any of the requirements will result in points deduction.

3.1 Data

You will obtain a data set and identify appropriate research questions. The choice of the dataset should be clearly motivated by your research questions (Section 3.1.1), and its properties (size, number of features and classes, skewness, linear separability, etc) should be taken into account when choosing the dataset and the methods.

The dataset should be complex enough in terms of the number of values and features to address the research questions. Small and simplistic data sets like those used in the example code (Iris, Titanic etc) are not considered complex enough unless the group comes up with an unusually creative research question and performs an extremely thorough and deep analysis. You can choose relatively simple datasets if you demonstrate that they are sufficient for an important research problem (for example, you have a dataset of patients with a rare disease) and you are still able to derive meaningful patterns and results from this data and explain them. Larger data sets with richer features, on the other hand, will naturally make it easier formulate interesting questions and support a variety of experiments. As a rule of thumb, we expect data sets to contain tens of thousands of instances to allow for robust training and evaluation. We also expect a certain complexity in the feature space (such as a large number of original or engineered features).

Datasets should be publicly available. Make sure accessing the data does not require special approvals that would take hours/days since this can delay your project. Your report must include the URL(s) of the dataset(s). The URL must be publicly accessible, and cannot lead to your own repository or site. You are not allowed to generate or extend existing datasets; in case of doubt we will check the date of the dataset publishing, and reject any datasets which seem to be AI-generated. Using AI to generate or extend dataset will be treated as academic misconduct. Your dataset(s) should be large and complex enough to address your research questions and support training and testing your chosen algorithms (i.e. "not enough data" excuse for getting poor and unconvincing results will not be accepted).

Suggestions: Possible datasets can be found at Kaggle (https://www.kaggle.com/datasets?fileType=csv), the UCI Machine Learning Repository (https://archive.ics.uci.edu/), or Hugging Face (https://huggingface.co/datasets).

Suggestions: While it is not a requirement, you are encouraged to use and integrate features from multiple datasets.

Comments: You will not be given more marks for the dataset size and complexity. Avoid spending too much time on computation or memory issues because the data is too big. If you already have experience with big data, parallelism, etc., then you can make use of those things, but you will NOT get a higher grade just based on that fact, and questions related to technical aspects of dealing with large data sizes or optimisation CANNOT be a research question. Of course, you CAN explore questions related to the effect of data size vs model complexity on the results. If the dataset you chose is too large, you may use a random subset of the data, explaining in the report why and how that was done.

3.1.1 Research questions

You will **identify a problem** to investigate in your project, explain why the project is interesting and why your chosen data set is appropriate. 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.

The general problem you are trying to address must be split into several **specific and well-defined research questions**.

- **Groups of 3 members are expected to define and address at least 2 research questions**
- **Groups of 4 members are expected to address at least 3 research questions**

Each research question must address a different aspect of the problem rather than propose different methods for the same task. For example, you two research questions such as "Is Naive Bayes an appropriate method for this data?" and "Is SVM an appropriate method for this data?" are *not* appropriate. Most research questions will involve their own model comparison.

Suggestions: You may use generative AI tools to assist you in developing ideas and research questions. However, the use of genAI must be clearly specified in the disclaimer of the report.

3.1.2 Understanding the data, feature selection and construction

You must apply at least ONE data preprocessing OR data analysis, although you are encouraged to use multiple if they enhance your analysis. Your chosen method(s) should align with your research question, and you should justify your selection in your report.

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, or analysis) is not sufficient and marks will be deducted.

Advice. 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 they will receive less marks than the more complex methods. You are advised to create new variables from the existing ones, perform feature selection and remove unnecessary features, handle outliers, combine features from multiple datasets, etc.

Suggestions. Some ideas for complex manipulations (for your reference only, DO NOT replicate them in your report):

- 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 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.

3.1.3 Machine Learning methods

For this assignment, at least ONE of the research questions should compare several machine learning models with a neural network. For instance, if you are performing classification, you should choose 3 classifiers, e.g., Naïve Bayes, Support Vector Machines and Decision Trees, and compare them to a neural network.

- **Groups of 3 should include at least 3 machine learning models and 1 neural network.**
- **Groups of 4 should compare at least 5 models, at least one of which is a neural network (for example, 3 classic ML models and 2 neural models with different architecture).**

Do not use the same algorithm with just different hyperparameters (e.g., SVM with C=1 and SVM with C=10). The algorithms need to be different.

Requirement. Choose algorithms that cover some spectrum from high-bias/low-variance to low-bias/high-variance. Compare them and explain the results of these models taking into account the properties of your data. Every algorithm should have at least one hyperparameter to be tuned, reported, and explained.

Requirement. All experiments should be properly validated. You can use reserved splits (train/dev/test) or cross-validation (including nested cross-validation for hyperparameter tuning). Each of the machine learning algorithms should follow the same validation approach, in order to fairly compare them in your report. Validation choices should make sense for your data type and size.

3.2 Report

The report should follow the structure of a research paper, as discussed in the guest lecture on Academic Writing. It should describe your approach and observations in the context of your chosen problem and research questions, both in data preparation, feature selection and engineering, and the machine learning algorithms you tried. We want to see evidence of your thought processes and reasoning for choosing one method over another and, in particular, **critical analysis of your results and discoveries**. The internal structure of well-known machine learning models should only be discussed if it is important for connecting the theory to your practical observations.

The report should include the following required content. Omitting any sections or not including the required information in it will be penalized:

- **Introduction**: a short description of the problem and the research questions addressed, their motivation and importance. While your motivation can be personal ("I chose this problem because I like soccer"), when writing up the report, provide reasons why you think it is an interesting or useful problem in general ("Prediction of soccer players behaviour in different weather conditions can provide insights to improve their training"). The specific research questions you address in the report must be clearly listed and linked to the general problem here, and later connected to your data, proposed methods, and analysis. Here or in Methods, you should also provide general information regarding the chosen dataset, and explain why this dataset is the most appropriate one to explore the research questions. This section sets the context for what follows and explains why the analysis was conducted.
- **Literature review**: a short summary of related literature, including the reference to your chosen data set (such as the original paper where it was published) and at least two additional relevant research papers of your choice. For each included paper it must be clear what it did and how it relates to your work.
- **Method**: Describe the dataset in terms of its properties (size, number of features and classes, skewness, linear separability, etc) and explicit link them to the research questions, the choice of models and analysis of the results. Include the statistics of data splits, or details of cross-validation, and explain the choice of validation setup, linking it to your data and models. The proposed/engineered/selected features must be motivated in the report by your intuitions regarding the problem and methods. Explain the reasons behind using particular ML models, tying them to your research questions. The range of chosen hyper-parameters and optimal value must be reported, and the effect of the hyper-parameters on the final result should be described and explained. The chosen evaluation methods should be motivated, i.e., explain why they are appropriate for your question, methods and/or data. When writing this section, you can assume that the reader is familiar with the technical terms.
- **Results**: Present the results, in terms of evaluation metric(s) and some illustrative examples. You should present the results using at least 3 metrics or evaluation methods such as learning curves, confusion matrix, precision, recall, accuracy, F1-score, root mean squared error, mean absolute error, etc. Include at least one table and diagram.
- **Discussion**: Contextualise** the system's behavior, based on the understanding from the subject materials as well as in the context of the research questions. Perform error analysis of the results, highlighting specific cases where some of the methods worked better or failed. Use visuals such as confusion matrices or error examples to support the analysis. Connect the observed behaviour to the properties of your models and data, and explain the limitations of the chosen methods accordingly. Show evidence of attempting to overcome such limitations (for example, if your intuition is that the data is not linearly separable, add an experiment with a non-linear model).
- **Conclusion**: Clearly explain how the research questions were answered. Summarise the main points of the paper and reiterate the key findings and recommendations.
- **Generative AI usage**: Declare any generative AI tools used for this project, and the reasons and scope/extent of their usage (see Section 4 below). This section does not count towards page limit.
- **Bibliography**, which includes references to the dataset and any related work you used in your project. For both bibliography and in-text citations, follow the style specified in the ACL template. This section does not count towards page limit.

** Contextualise implies that we are more interested in seeing evidence of you having thought about the task, and determined *reasons* for the relative performance of different methods, rather than the raw scores of the

different methods you select. This is not to say that you should ignore the relative performance of different runs over the data, but rather that you should think beyond simple numbers to the reasons that underlie them.

4 Assignment Policies

Terms of Data Use

You must use a publicly available dataset and properly reference it (including a working link). All dataset links will be checked for validity. If a dataset was generated using AI tools, this will be treated as academic misconduct. This will apply even if AI-generated datasets are uploaded to a public repository Huggingface. To prevent such misconduct, we will not accept ay datasets released or updated on or later than Monday Sep 15th (Assignment release date).

Changes/Updates to the Project Specifications

We will use LMS announcements for any large-scale changes (hopefully none!) and Ed for small clarifications. Any addendums made to the Project specifications via LMS will supersede information contained in this version of the specifications.

Late Submission Policy

We allow **no extensions or late submissions** considering that the assignment is a group project. Submission will close strictly at **5 PM on Friday October 22th**. Students who are eligible for **special consideration** (e.g., with an APP) please email Lea Frermann (`lea.frermann@unimelb.edu.au>`) and a solution will be sought.

Academic Misconduct

The reuse of ideas between groups will be considered cheating. We highly recommend to (re)take the academic honesty training module in this subject's LMS. We will be checking submissions for originality and will invoke the University's Academic Misconduct policy[1] where inappropriate levels of collusion or plagiarism are deemed to have taken place.

Use of generative AI

Content produced by generative AI (including, but not limited to, ChatGPT) is *not* your own work, and submitting such content will be treated as a case of academic misconduct, in line with the University's policy. In particular, generative AI must not be used to generate any part of your report, including using AI to write paragraphs or sections based on prompts. Similarly, translating the report or any of its parts using genAI tools is prohibited and will be treated as academic misconduct. As explained above, the same policy applies to generating datasets.

[1] http://academichonesty.unimelb.edu.au/policy.html

You are allowed to use generative AI for a) idea and research question development; b) paraphrasing or translating individual words or short phrases less than 1 sentence long.

All usage of AI must be declared in the paper after Conclusions in a separate section (such as *Generative AI usage*). Please see the guidelines here: https://students.unimelb.edu.au/academic-skills/resources/academic-integrity/acknowledging-AI-tools-and-technologies**. This section does NOT count towards the page limit.**