# Hong Kong University of Science and Technology
# COMP 4211: Machine Learning
# Spring 2024

## Group Project

Proposal due: 15 April 2024, Monday, 11:59pm
Report, code, video, and slides due: 6 May 2023, Monday, 11:59pm

## Guidelines

- Each project group should have two students from the ***same*** lecture section. You may form your own project group. We may also form a group for you if you need help.
- Project submission should be done via Canvas with two deadlines:
  - Proposal: The proposal file named `proposal.pdf` should be no more than two A4 pages in length.
  - Report, code, video, and slides: The zipped submission should contain the report `report.pdf`, compressed source code `code.zip`, and presentation slides `slides.pdf` or `slides.pptx` (with the hyperlink of the video presentation included in the first slide).
- The project will be counted towards 20% of the final course grade, of which the proposal accounts for 2%, report and source code account for 13%, and video presentation and slides account for 5%.
- As always, please use Piazza to ask TAs and professors if you have any questions.
- No generative artificial intelligence (GenAI) tools for code and text generation are allowed.
  - Prior approval from the course instructor is needed if you want to use any other GenAI tool for your project.
  - If approved, it should be stated clearly in the project report with detailed explanation on how the GenAI tool is used.
- Please refer to the regulations for student conduct and academic integrity on this webpage: `https://registry.hkust.edu.hk/resource-library/academic-standards`.

# 1 Preamble

The objective of this project is to practise the hands-on skills needed for solving more realistic machine learning tasks through pursuing a proposed study that involves one or more machine learning topics studied in this course.

Unlike the programming assignments and problem set, this project is intended to be more open-ended like many other course projects or final year projects. As such, much room is left for you to explore. Consequently, there will only be grading guidelines but not a detailed marking scheme.

The project is expected to be more substantial and hence is a group project. More information about group formation will be announced in Piazza later if you need help.

Note that this project should not be used for earning credits in a different course to avoid double-dipping.

# 2 Project Ideas and Considerations

Your proposed project must involve one or more machine learning topics covered in this course. One example is to apply a machine learning model learned in this course to solve a real-world problem. While you do not have to strive for developing a new model or a new algorithm, novelty of the application in terms of applying the model to solve it will be a criterion in the grading guidelines to be detailed later.

For inspiration, you may take a look at the Kaggle website (`https://www.kaggle.com`) and other online resources. Publicly available datasets may be used for the project. Note that sometimes a dataset originally used for one task may be used in a very different way for another task that has not been studied by others before. Try to think out of the box to explore novel applications of machine learning. It may even help you start a business in the future.

In case you plan to use some more advanced machine learning methods not covered in the course, please make sure that you also include the related methods covered in the course as baselines for comparison. Among other things, including the baselines will help you justify using more advanced methods.

One important thing to keep in mind is to ensure that the computing resources needed for the project are not exceedingly high. For example, you need to judge whether it is feasible to carry out the proposed project using only a subset of a large dataset, instead of the full set.

# 3 Project Proposal

You are expected to clearly describe at least the following in the proposal:

- Background information about the application, preferably explained and articulated using a simplified example for illustration that can appeal to a general audience

- Formulation of the application as a machine learning problem, which includes but is not limited to describing the problem type as well as the model input and output

- Description of related applications of machine learning in the literature, if any, with citation of relevant references

- Description of the dataset, which includes but is not limited to its size and features
- Expected computing resources needed, within the limit of what will be available to your group

The proposal is worth 2% of the final course grade. You will get full marks for this part if all the required aspects above are presented clearly.

## 4 Project Report and Source Code

The report should cover at least the following aspects of the project:

- Project title
- Students with full names, student IDs, and HKUST email addresses
- Description of the dataset and any preprocessing
- Description of the machine learning task(s) performed on the dataset
- Machine learning method(s) used for solving the task(s)
- Experiments and results
- Division of labor in teamwork

The project report should be self-contained in the sense that readers can understand what you do just by reading the report without having to refer to the proposal. As such, some materials in the proposal are expected to be included in the report as well.

In case it is deemed necessary for you to deviate from what you plan to do as described in the proposal, you should provide strong justification for doing so in the report. While a slight deviation from the proposed work is generally fine, a significant change should only be made as the last resort with sound justification.

You should state clearly the division of labor between the two group members by listing the main duties and contributions of each member. In addition, the overall contribution of each member to the project should also be given in percentage (e.g., 55% by A and 45% by B). You should try your best to ensure that the workload is shared evenly (i.e., 50% each). Grading will be done individually according to the workload distribution.

All the source code that you have written and used for this project should be submitted for grading. In case your code is modified from another source, you should acknowledge it clearly in your report and point out which parts are yours. Failure to do so is considered plagiarism.

Data files should not be submitted to keep the submission file size small.

You should name the report as `report.pdf` and the compressed source code as `code.zip`.

The report and source code are worth 13% of the final course grade. The breakdown is as follows:

- Description of the dataset and any preprocessing (15 points)
- Description of the machine learning task(s) performed on the dataset (10 points)
- Description of the hardware and software computing environment, machine learning method(s), and parameter settings (10 points)
- Good programming practices in the source code (10 points)
- Novelty of the application (10 points)
- Description of the experiments (20 points)
- Visualization and discussion of the results obtained (20 points)
- Proper division of labor (5 points)

Suppose you want to reduce the input feature dimensionality of your dataset first before using it for a classification task. The two main machine learning tasks involved are then dimensionality reduction and classification, with the former being an unsupervised learning task while the latter a supervised learning task. The tasks should be described in detail, involving but not limited to description of the input and output of the model for each task. You then report the machine learning methods used in your experiments for solving each of these two machine learning tasks, possibly involving performance comparison of different methods for each task. When you describe your experiments, you should first describe the experiment setup before presenting the results obtained with visualization and discussion. It would also help to discuss possible extensions that may be able to bring about improvement.

The project report should be self-contained even without referring to the code. An important general criterion is clarity, to the extent that others can replicate your experiments based on the information provided in the report.

It is noted that achieving better performance for a model often requires extensive hyperparameter tuning. To save your time and the need for substantial computing resources, the grading criteria will not put too much emphasis on achieving superior performance as long as it is reasonably good. As said above, more emphasis should instead be put on novelty of the application itself. Moreover, you may regard this as a proof-of-concept study and focus more on analyzing the pros and cons of a certain model design for the dataset used.

# 5 Video Presentation and Slides

You are required to prepare an oral presentation of your project in the form of a video using informative slides that summarize the key aspects of the project. The video should be 15 minutes in length, with deviation by no more than 10% acceptable. The faces of the group members should be visible during the presentation.

Note that the video is not a movie for entertainment or an advertisement. Instead, it is for a technical presentation of your project. You should pay attention to both the technical content and the quality of your video.

When your video is ready, upload it to YouTube as an ‘unlisted’ (not ‘private’ or ‘public’) video and include its hyperlink clearly in the first slide. The video should be ready by the time you submit the slides and no change should be made to it after the deadline. Note that it takes time to upload a video of 15 minutes long to YouTube, so you should manage your time well instead of rushing towards the end. No request for extension of deadline will be entertained for failing to complete video upload by the deadline.

The target audience of your presentation should be students immediately after taking COMP 4211, i.e., students who have learned the machine learning topics in the course. In other words, in your presentation, you may refer to concepts and techniques covered in the course without having to review them again.

You are expected to present informative slides that summarize the key aspects of the project. If some of your slides are modified from another source, you are expected to acknowledge it clearly. Failure to do so is considered plagiarism. Note that solely including the source in the list of references at the end of your presentation is necessary but not sufficient.

You should name the file as `slides.pdf` or `slides.pptx` in your submission.

The video presentation and slides are worth 5% of the final course grade. Here is the breakdown:

- Good time management in the presentation (20 points)
- Clarity and good organization of content in the slides (30 points)
- Clarity of the oral presentation (30 points)
- Effective use of examples and visual aids (20 points)

## 6 Computing Facilities

You should estimate the computational demand of your proposed project. In case the computing resources provided by the basic Colab plan cannot meet your need, you may consider a short-term subscription of the Colab Pro/Pro+ plan or other cloud computing services. If applicable, you may use a pre-trained model as the starting point for the training of your model. Not only does it shorten the training time needed, but it also allows the model to be (further) trained using a smaller training set.

## 7 Submission

All assessment components of the project must be submitted electronically in the Canvas course site, with two separate deadlines as listed on the first page.

Only one member of each project group will submit all the assessment components on behalf of the group, but the names of both members should be listed clearly in all the assessment components.

When multiple versions with the same filename are submitted, only the latest version according to the timestamp will be used for grading. Files not adhering to the naming convention above will be ignored.

Late submission will be accepted but with penalty. The late penalty is deduction of one point (out of a maximum of 100 points) for every hour late after 11:59pm with no more than two days (48 hours). Being late for a fraction of an hour is considered a full hour. For example, two points will be deducted if the submission time is 01:23:34.

## 8 Grading

This project is intended to be more open-ended. As such, much room is left for you to explore according to your interests.

Grading will be based on rubrics with five levels of achievement (excellent, good, satisfactory, unsatisfactory, poor) for each of the assessment items.