# DATA7703 - Machine Learning for Data Scientists
# Group Project

Semester 2, 2023

**Topic**
a topic of your choice on machine learning
**Group size**
4 to 5
**Deliverables**

| | | |
|---|---|---|
| proposal | 5% | due Week 9 (submit a PDF file) |
| seminar | 10% | due Week 13 (submit slides and give a talk) |
| report | 25% | due start of Exam Period (submit a PDF report, and code if applicable) |

## Topics

The project is where you showcase your capability to pool your efforts together to apply machine learning to solve a substantial data science problem. You may choose to work on an application project, an algorithm/method project, a review project, or any type of project that you are keen to work on, as long as the scope is substantial enough to demonstrate strong understanding of machine learning.

*Application.* An application project mainly involves developing a machine learning solution to a practical problem. For example, you may apply techniques like SVMs, random forests, or neural nets to a challenging practical problem.

Your analysis may involve answering some or all of the following questions below.

- (Exploratory analysis) Are the covariates correlated? How are the covariates distributed?
- (Model diagnostics) Is/Are the fitted model/models appropriate for the dataset (e.g. whether there are evidence that the model assumptions may be inappropriate)? Can we use them to draw conclusions?
- (Diagnostic analysis) What are the important factors/covariates?
- (Predictive analysis) What does the model predict on new data/situation? Does the model generalize well on new data? Can we interpret the model's prediction?
- (Prescriptive analysis) What actions should be taken based on the model?

In general, you do not need to answer all these questions. For example, for some applications, successfully developing a machine learning system with good predictive performance is already a very substantial project.

While a data science project typically puts a heavy emphasis on many other aspects of data science — such as data collection, data cleaning, visualization of data and analysis — and your project is most likely to include some of these, your project's most significant component should be a showcase of how you can skilfully apply machine learning techniques to solve an

interesting problem. Typically, this implies that you need to choose a challenging application that cannot be solved by rudimentary techniques like linear regression.

In addition, you are likely to explore several techniques before you reach a satisfactory solution. What is satisfactory may depend on the application, but usually you are looking for things like simplicity, good predictive performance, interpretability, robustness. You will need to justify how you reach your solution, and why you consider that as acceptable.

*Algorithm/method.* An algorithm project requires developing a new algorithm for a machine learning task. It may be about modifying some existing models/algorithms and see whether that works, but it can also be about proposing a highly novel approach. As another example, you can also perform extensive experiments to study how different methods/models compare with each other on a range of datasets (e.g. extensive comparison of different classification methods on datasets of different characteristics).

*Review.* A review project need to present a thorough and comprehensive discussion on an advanced statistics topic. It can be a synthesis project in which you extensively and intensively read about a topic of interest, and then write down what you learned in a self-contained coherent report.

**Data sources**. Here are some sources of publicly available data that you may find helpful for your project. If you already have a problem that you want to solve by doing some statistical analysis, you can go to these data sources and search for the data you need. If you do not have a specific idea yet, it will be helpful to browse through some of the datasets to see whether you find something interesting.

- Kaggle datasets: `https://www.kaggle.com/datasets`.
- Some government datasets: `https://www.data.gov/`, `https://data.gov.au/`, `https://data.gov.sg/`.
- UCI machine learning datasets: `https://archive.ics.uci.edu/ml/datasets.php`.
- FiveThirtyEight (interactive sports and news site); `https://data.fivethirtyeight.com/`.
- BuzzFeedNews (America Internet media, news and entertainment company): `https://github.com/BuzzFeedNews`.
- Socrata (cleaned open source data sources ranging from government, business, and education data sets): `https://opendata.socrata.com/`.

## Deliverables

**Proposal**. The proposal file should be a PDF file, limited to 2 pages, with font size of at least 11pt. The proposal should have the following components.

- Title and topic: write down the title of your project and explain the proposed problem.
- Significance: explain why the problem is important.

- Feasibility: justify that the project is feasible by explaining what has been done, what the planned work will be and how it will be carried out. If data is needed for your project, make sure that you have found some, as sometimes data does not come as easily as we expect.

The proposal will be assessed based on its clarity and conciseness, significance, feasibility, and novelty.

**Seminar**. Each group will give a talk to showcase their project. The slides for the talk can be in any commonly used format (such as .ppt file or .pdf file). The length of each talk will be announced closer to the seminar date, but most likely will be around 10 minutes for each talk.

The talk should clearly convey to the audience the following aspects of the projects: the problem, significance of the problem, steps taken to solve the problem, main findings, limitations of the study, and conclusion.

The talk will be assessed based on whether the above points are clearly explained, and whether the claims are compellingly supported with adequate amount of technical details (See Tips section for some examples). While it is understood that a short talk cannot provide all the technical details needed for reproducibility, a lack of technical details will make your talk look like a show of opinions, rather than a scientific presentation.

**Report**. This is the most significant deliverable of the project.

There is no page limit on the report, but your report is expected to cover the following aspects of your project: the problem, significance of the problem, steps taken to solve the problem, main findings, limitations of the study, and conclusion.

Your report will be assessed based on the following criteria.

- Clarity and conciseness: no spelling, grammar, and punctuation mistakes; content is well-organized; notations and technical statements are clear; writing is straight to the point.
- Correctness: the arguments and technical details should be sound and correct.
- Reproducibility: the report should include sufficient details so that a reader can follow the report to reproduce the presented results.
- Significance: the project can be significant in various ways, including the choice of an important societal or scientific problem, the amount of effort or creativity required to develop the reported solution, the thoroughness of the study, the potential impact of the findings. You should highlight in your report what you believe to be significant.
- Novelty: demonstration of creativity by proposing new ways of formulating an application as a machine learning problem, finding out limitations of algorithms that have not been observed for the application, developing new methods to interpret the model for the application... You should highlight in your report what you believe to be novel.

On a blank page immediately after the title page, indicate whether all your group members consent for your report to be used as a teaching resource by including one of the statement

below.

*We give consent for this to be used as a teaching resource.*
*We DO NOT give consent for this to be used as a teaching resource.*

For a project involving implementation, you should submit your code separately as a zip file on Blackboard.

## Tips

A good project need to demonstrate good understanding of the problem, your skills in crafting a solution, rigorous and thorough evaluation of your solution, and effective communication of the significant process and findings via a presentation and a report.

The following are some tips that you may find helpful, particularly if you are working on an application or algorithm project.

- Problem and significance
  - State the problem statement by setting the objectives to be achieved, with an explanation of relevant background or context needed for understanding it.
  - The problem's significance can be demonstrated by explaining how it is related to other problems, its applications, with citations to related references.
- Solution
  - Provide model and training details. It need to be detailed enough so that your approach can be reproduced and correctness can be verified. A technical project shouldn't be assessed based on whether I trust you.

    For example, instead of just saying "we train a random forest", also provide important details such as number of trees, maximum tree depth, number of random features considered at each split... In particular, if you are using the implementation in a library, and you are not using the default parameters, make sure you report the values that you choose.
  - Highlight important tricks that you have used, e.g. data augmentation, design of domain-specific loss, balancing dataset, ...
- Findings
  - Explain the setup for evaluating your system clearly. In particular, describe the datasets and its characteristics, and the performance measure.
  - Plots of learning curves (train/test performance against iteration/epochs) are very useful for identifying obvious convergence issues.
  - An overall performance of your system should be reported on a test set.

    Many problems have well-defined and commonly used performance measures, e.g., we use measures like accuracy/precision/recall/F-measures for classification problems, and MSE for regression problems. Your system's test set performance should

be included to reflect the solution's overall performance (including training set performance helps as well).

In cases where there is no commonly used numerical performance measure, evaluation on some well-designed cases will be indicative of the system's performance.

- A more in-depth analysis on the system's performance is to analyze when the system does well and does poorly.

  For example, for classification, confusion matrix allows us to which classes are easy and which are hard, and common errors. Examining examples when the system works and fails can also provide insights on the system's strength and weakness.

- Limitations
  - If a satisfactory solution has not been found, this will be the place to discuss practical difficulties, and what you would have explored for making things work if time permits.
  - Clarifying limitations in your solution is very helpful for readers to better interpret your results.
  - Any other limitation of the study can be discussed here as well. The main purpose is to help readers to better understand your approach and results. It also shows that you have holistic understanding of your study.
- Conclusion
  - This can include a summary of significant findings, e.g. what you found to be most helpful for getting good performance, what makes the problem difficult if a good solution has not been found. This should help readers to get a good idea of what this project has achieved quickly.
  - This can also include suggestions based on your findings.