**DATA4001 Final Project Information**

One of the module's main goals is to prepare you to apply predictive modelling and ML algorithms to real-world tasks. The final project is intended to start you in these directions.

**Project Topics**

Your first task is to pick a project topic. If you're looking for project ideas, contact me, and I'd be happy to brainstorm and suggest some project ideas. In the meantime, here are some [suggestions](https://docs.google.com/document/d/1Cv0zs-B4kzMdKVzqwaKBMtr8TNAA1S68kZbwwv1GOUM/) that might also help.

Students can do an **Application project**. Pick an application that interests you, and explore how best to apply learning algorithms to solve it.

Many fantastic class projects come from students picking either an application area that they're interested in or picking some subfield of machine learning that they want to explore more. So, pick something that you can get excited and passionate about! Be brave rather than timid, and do feel free to propose ambitious things that you're excited about. (Just be sure to ask us for help if you're uncertain how to best get started.)

Once you have identified a topic of interest, it can be useful to look up existing research on relevant topics by searching related keywords on an academic search engine such as [http://scholar.google.com](https://www.google.com/url?q=http://scholar.google.com/&sa=D&ust=1507334628896000&usg=AFQjCNFWyWHMR3z5lxXXIt6hDiY6edS0DQ). Another important aspect of designing your project is to identify one or several datasets suitable for your topic of interest. If that data needs considerable pre-processing to suit your task, or that you intend to collect the needed data yourself, keep in mind that this is only one part of the expected project work, but can often take considerable time. We still expect a solid methodology and discussion of results, so pace your project accordingly.

Notes on a few specific types of projects:

* **Deep learning projects**: This module discusses many other concepts besides deep learning, we ask that if you decide to work on a deep learning project, please make sure that you use other material you learned in the class as well. For example, you might set up logistic regression and SVM baselines, or do some data analysis using the unsupervised methods covered in class. We may grade these projects using different criteria to make sure that grading is fair for students who have not had exposure to DL before. Finally, training deep learning models can be very time consuming, so make sure you have the necessary computer.
* **Preprocessed datasets**: While we don't want you to have to spend much time collecting raw data, the process of inspecting and visualizing the data, trying out different types of preprocessing, and doing error analysis is often an important part of machine learning. Hence if you choose to use preprepared datasets (e.g. from Kaggle, the UCI machine learning repository, etc.) we encourage you to do some data exploration and analysis to get familiar with the problem.
* **Replicating results**: Replicating the results in a paper can be a good way to learn. However, we ask that instead of just replicating a paper, also try using the technique on another application, or do some analysis of how each component of the model contributes to the final performance.

**Project Parts: Proposal, Poster and Final Report**

This section contains detailed instructions for the different parts of your project.

**Submission**: MS-Word Format

**Evaluation**

We will not be disclosing the breakdown of the 40% that the project is worth amongst the different parts, but the poster and final report will combine to be the majority of the grade. Projects will be evaluated based on:

* The technical quality of the work. (I.e., Does the technical material make sense? Are things tried reasonable? Are the proposed algorithms or applications clever and interesting? Do the authors convey novel insight about the problem and/or algorithms?)
* Significance. (Did the authors choose an interesting or a “real" problem to work on, or only a small “toy" problem? Is this work likely to be useful and/or have an impact?)
* The novelty of the work. (Is this project applying a common technique to a well-studied problem, or is the problem or method relatively unexplored?)

In order to highlight these components, it is important you present a solid discussion regarding the learnings from the development of your method, and summarizing how your work compares to existing approaches.

**Project Proposals 20% out of 40%**

In the project proposal, you'll pick a project idea to work on early and receive feedback from the lecturer. If your proposed project will be done jointly with a different class' project, you should obtain approval from the other instructor and approval from the module's instructor.  Please discuss with me if you would like to do a joint project.

In the proposal, below your project title, include the project category. The category can be one of:

|  |  |
| --- | --- |
| * Athletics/Sports & Sensing Devices * Audio & Music * Computer Vision * Finance & Commerce * General Machine Learning | * Retail/ Supply Chain * Healthcare * Sales/Marketing * Natural Language |

|  |  |
| --- | --- |
|  |  |
| Format | Your proposal should be a MS WOrd document, giving the title of the project, the project category, the full names of all of your team members,  and **a 300-500 word description of what you plan to do.**  Your project proposal should include the following information:   * **Motivation**: What problem are you tackling? Is this an application or a theoretical result? * **Method**: What machine learning techniques are you planning to apply or improve upon? * **Intended experiments**: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?   Presenting pointers to one relevant dataset and one example of prior research on the topic is a valuable (optional) addition. |
| Grading | The project proposal is mainly intended to make sure you decide on a project topic and get feedback early. As long as your proposal follows the instructions above and the project seems to have been thought out with a reasonable plan, you should do well on the proposal. |

**Final Writeup+ poster 80% out of 40%**

I know that most students work very hard on the final projects, and so we are extremely careful to give each writeup ample attention and read and try very hard to understand everything you describe in it.

|  |  |
| --- | --- |
| Format | Final project writeups can be at most 5 pages long (including appendices and figures). We will allow for extra pages containing only references. If you did this work in collaboration with someone else, or if someone else had advised you on this work, your write-up must fully acknowledge their contributions. For shared projects, we also require that you submit the final report from the class you're sharing the project with.  Here are more detailed guidelines with a rough outline of what we expect to see in the final report: [final-report-guidelines.pdf](http://cs229.stanford.edu/final-report-guidelines.pdf). |
| Contributions | Please include a section that describes what each team member worked on and contributed to the project. If you have any concerns working with one of your project teammates, please create a private Brightspace post. We may reach out and factor in contributions and evaluations when assigning project grades. |
| Code | Please include a zip file or preferably a link to a Github repository with the code for your final project. You do not have to include the data or additional libraries (so if you submit a zip file, it should not exceed 5MB). |
| Grading | The final report will be judged based on the clarity of the report, the relevance of the project to topics taught in the module, the novelty of the problem, and the technical quality and significance of the work. |

There will be **no late days** for the final report.

**Project FAQs**

1. What are the deliverables as part of the DATA4001 term project?

 The project has four deliverables:

1. Proposal
2. Final report+ Poster

2. Should the final project use only the methods taught in the classroom?

No, we don't restrict you to only use methods/topics/problems taught in class. That said, you can always consult the instructor if you are unsure about any method or problem statement.

3. Is it okay to use a dataset that is not public?

We don't mind you using a dataset that is not public, as long as you have the required permissions to use it. We don't require you to share the dataset either as long as you can accurately describe it in the Final Report.

4. What fraction of the final grade is the project?

The term project is 40% of the final grade.

5. can we use some Machine Learning libraries such as scikit-learn (Python) and CARET (R) or are we expected to implement them from scratch?

You can use any library for the project.

6. Will we be provided any cloud compute resource credit?

Check out Google Colab for free GPU resources.

7.  Are we required to use Python for the project?

Any programming language or tool is allowed for the project. I expect to use rapidminner.

**Final Report**

**Deadline:** May 17th

The final report should contain a comprehensive account of your project. We expect the

report to be thorough, yet concise. Broadly, we will be looking for the following:

• Good motivation for the project and an explanation of the problem statement

• A description of the data

• Any hyperparameter and architecture choices that were explored (Predictive

modelling)

• Presentation of results

• Analysis of results

• Any insights, implications, recommendations and discussions relevant to the project

• References

**Format**

Final project writeups can be **at most 5 pages** long (including appendices and

figures). We will allow for extra pages containing only references.

You are strongly encouraged to use the below format.

If you are not using this format, make sure to include all sections given in the format.

**Contributions**

**(if any)**

Please include a section that describes what each team member worked on and

contributed to the project.

**Sources file** Please include a link or attach **Models, dashboards, Process Analytics model**

with your final project. You do not have to include the data or additional libraries.

**Grading**

The final report will be judged based off of the clarity of the report, the relevance of the project to topics taught module, the novelty of the problem, and the technical quality and significance of the work.