COMP 381: Final Project

Overview: There are two possible options for this project. You can find a dataset and apply some of the machine learning algorithms and evaluation methods that we have covered to the dataset, or you can implement one or more of those algorithms from scratch.

Option A. Applying Machine Learning

The final project should demonstrate your in-depth understanding of at least **three** machine learning algorithms, including how to use the algorithms in Python / sklearn, how to apply them to a specific dataset, and how to evaluate performance of the algorithms. You have a great deal of flexibility in your project, as long as it meets these requirements and is clearly worthy of two+ weeks of work (i.e. it is not thrown together at the last minute). If you are not sure whether your topic is appropriate, come talk with me.

This is an individual project (no partners or teams). However, feel free to discuss your project with your classmates.

Details: Find a dataset (see links below) and apply at least three algorithms to the dataset, using Python and sklearn. Your results and report should answer at least the following questions:

- What is your dataset? Where did you find it, and what does it represent? How many features are there? How many observations?
- Do you have dedicated training and test sets? If not, are you doing cross-validation?
- Are you doing classification, regression, or both?
- Is it a supervised or unsupervised task?
- What are the specific algorithms you are using?
- Are you using the full set of features, or do you carry out feature subset selection?
- Do you use regularization methods?
- What are your evaluation metrics?
- What are the most interesting findings?
- Provide any relevant visualizations / graphs.

Datasets: I will be posting on Blackboard a large number of links to datasets. Feel free to post a link to a dataset you found if you think other students will find it useful as well. Here are three places to find data, to get you started in your search:

- Kaggle - https://www.kaggle.com/datasets
- UCI ML Repository - https://archive.ics.uci.edu/ml/datasets.html

- Government of Canada - http://open.canada.ca/data/en/dataset
- Figure Eight: https://www.figure-eight.com/data-for-everyone/

Deliverables:

- Submit a Jupyter notebook containing your code and textual content that comprehensively answers the questions listed above.
- Also submit your dataset, or a link to the dataset, or a sample of the dataset.

Option B. Implementing Machine Learning Algorithms

The second option is to implement from scratch one or more of the algorithms we have covered. Again, this must be a suitable amount of work for two+ weeks.

If you choose this option, you **must** create a CIS GitLab project and add me to it so that I can monitor your progress while you are implementing the algorithm. If you only submit final code and I have not been able to see the development of it, you will not get full credit.

https://cisgitlab.ufv.ca/users/sign_in

Deliverables:

- Submit your code or a link to your code on GitLab.
- Submit any instructions needed to run your code.
- Submit a 1-2 page report summarizing what you did, and any surprises or difficulties, etc.