Project Report Guidelines

# Introduction

* Project description
* Problem statement

# Supervised Learning

## Motivation (5 points)

Briefly state the nature of your work and why you chose it. (b) What specific question, goal, or task did you try to address related to structure in the data?

## Data Source (5 points)

Describe the properties of the dataset (or data API service) you used. Be specific. Your information at a minimum should include but not be limited to:

* where the datasets or API resource is located,
* what formats they returned/used,
* what were the important variables contained in them,
* how many records you used or retrieved (if using an API), and
* what time periods they covered (if there is a time element)

## Methods & Evaluation (20 points)

* Provide a correct and comprehensive evaluation, analyzing the effectiveness of both your methods, and your feature representations.
* What important tradeoffs can you identify?
* How sensitive are your results to choice of parameters, features, or other varying solution elements?

## Failure analysis (10 points)

* Select examples where prediction failed, and analyse why.

# Unsupervised Learning (40 points total)

## Motivation (5 points)

Briefly state the nature of your work and why you chose it. (b) What specific question, goal, or task did you try to address related to structure in the data (e.g. the clusters you found)?

## Data Source (5 points)

Describe the properties of the dataset (or data API service) you used. Be specific. Your information at a minimum should include but not be limited to:

* where the datasets or API resource is located,
* what formats they returned/used,
* what were the important variables contained in them,
* how many records you used or retrieved (if using an API), and
* what time periods they covered (if there is a time element)

For example, if you downloaded data or used API services, you should state the specific URLs to those files or resources in a way that is trivial for the instructor to retrieve them if needed.

## Unsupervised Learning Methods (20 points):

Briefly describe the workflow of your source code, the learning methods you used, and the feature representations you chose.

* How did you tune parameters?
* What challenges did you encounter and how did you solve them?

## Unsupervised Evaluation (10 points)

* What interesting relationships or insights did you get from your analysis?
* What didn't work, and why?
* To summarize your findings, include at least two visualizations (chart, plot, tag cloud, map or other graphic) that summarize your analysis.

# Discussion (20 points)

* What did you learn from doing Part A? What surprised you about your results? How could you extend your solution with more time/resources?
* What did you learn from doing Part B? What surprised you about your results? How could you extend your solution with more time/resources?
* What ethical issues could arise in providing a solution to Part A, and how could you address them?
* What ethical issues could arise in providing a solution to Part B, and how could you address them?

# Statement of Work (0 points, but required)

* You must include a statement that describes the contribution that each team member made to the project.

**\*\*Suggestions from the Instructors\*\***

1. Your report should not just describe your methods - it should justify your choice of them. This applies to important choices like the learning framework/choice of model, individual features or feature types, overall feature representation, and evaluation methods. For example, if you decide to use two families of prediction model that are similar, can you justify why? (Perhaps because you’re exploring a particular hypothesis or question.)
2. Your report should ultimately give at least some insight into why particular approaches were successful or failed -- either by themselves or compared to each other. At a minimum, you should at least focus on your failure analysis to characterize the types of failures that future work might try to address.
3. As you’re planning your remaining time, I recommend prioritizing insight into features ahead of insight into models - although of course ultimately it will depend on your particular project, and ideally you should include at least some analysis of model differences. If you want to argue that in your application, model choices contribute more to gains than feature choice, by all means do that...
4. In the discussion section, include at least a few sentences that identify important potential tradeoffs. You don’t necessarily have to show evidence for them, just note that they exist and why.
5. When discussing potential ethical issues, it’s not just a question of whether there may be particular privacy or ethical issues in the specific dataset or method: if that’s less of a concern, you should focus on the more general question of potential ethical issues around \*deploying\* a machine learning artifact for your task, i.e. the use of that supervised or unsupervised learning task in making policy/judicial/educational/health and public safety decisions, etc.