**Final Project Guidelines**

**Choosing A Topic**

In this project, you have two options for the general route you can take:

1. You may identify a set of data and corresponding analysis that has not yet been performed. This can be more exploratory in nature. You will implement this analysis with testing and documentation. In your final report, you should discuss whether/how the type of data influenced the analysis that was possible, the findings and limitations of your analysis, and what might be ways to validate your findings. To be clear, I am most interested in seeing a well thought out approach in this option. If you take on an interesting project with a reasonable approach and things don’t work out, that is ok. You can receive a perfect grade on this even if the project does not work out as intended or hoped.
2. As a second option, you may select a paper that implements a computational analysis. You will reimplement this analysis with thorough testing and documentation. In your final report, you should include a discussion about any elements that presented a challenge during your reimplementation, or discrepancies you found. Lastly, you should identify a creative extension of the analysis and implement it. Note that I prefer you at least consider option 1. However, option 2 is also acceptable and this choice will not impact the grading of your project.

I will take a broad view of what a reasonable project is here in terms of both question an approach. While this is a physics class, the project you chose doesn’t need to be a physics project. Anything that raises a question that is interesting to you (and approved my me) is fair game. Further, you should not feel restricted to using only the tools discussed in this course. We are only scratching the surface and you should feel free to use what ever machine learning tools that are most appropriate for your problem. The point is for to take this as an opportunity to think through a small project from top to bottom, identify data and models to work with, and implement them. An integral part of this project will also be your presentation of results in the form of a final written report and class presentation. To ensure you and I are on the same page regarding your project in advance, you will submit a short proposal in advance.

If you are looking for ideas, I can recommend a few resources with interesting data sets where you may be able to find something of interest to you.

UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/index.php>

Kaggle: <https://www.kaggle.com/datasets>

You can of course choose your project, no matter where the data / idea comes from.

**Proposal (DUE MARCH 24)**

The proposal should be less than two pages and describe the following items:

* What topic you plan to study and why the topic you chose is interesting
* What data is available for you to use in this project.
* What overall approach do you plan to take for the project and why
* Demonstrate that your project can be finished within a month
* Estimate the difficulty of your project

**I am available to discuss your ideas, and if in doubt, you should discuss your idea with me prior to submitting your proposal.** By the time you submit your proposal you should ensure that you have access to the requisite data and have an idea for how you will carry out the project. We are happy to help you with this.

This proposal is due on March 24. Again, I encourage you to run anything by me in advance if you have any questions about suitability. The first round of presentations will be on April 19, so this leaves you 3-4 weeks to complete what you propose.

**Final Report**

Your final report should be less than 1500 words and describe the following items:

* Introduction/Motivation: Why the topic you choose is interesting, and whether similar work has been done by others (novelty will not affect your grade, but it is always good to know if other people are doing the same thing)
* Problem definition: How to formulate your topic/question into a data analytics or modeling problem.
* Methods
  + Description of the algorithm(s) approaches you employed or designed
  + The software package you chose or your own implementation of the algorithm
  + How to use the software package or your own code
* Results
  + Quantitative evaluation of your method
  + Charting and/or visualization of your results
  + How your proposed approach has solved your question

Accompanying your final report should be a github repository of your analysis. This should include all the data and computational files to produce your results. Included in this should be a jupyter notebook that documents your results with both markup text, documented code, and results. Note that you can use more than one file to generate results. For example, you can produce utility .py files as necessary. But I want to see one notebook that brings all of the results together. Any result that is discussed in your report should be generated in this notebook. If necessary (for example if you use a non-standard library that needs to be installed), please include a sufficiently detailed README file in your repo.

**Final Presentation**

Final presentations will occur in the last two weeks of class in the allotted class period.

* 10 minutes for each presentation.
* An additional (up to) 5 minutes for questions and changeover.
* Talk should cover the same basic bulleted sections as in the final report listed above.

**Grading**

Your project will be evaluated

* Proposal (15%)
* Final Report (40%)
* Final Presentation (30%)
* Code and Implementation (15%)