Math 475/575
Project #3:
due End of Day, November 18

Overall Description: In this Project, you will practice the kind of "higher-level" project planning that is used in Machine Learning project. You will generate

1. choose a data set
2. perform a very preliminary initial analysis of a data set
3. create a schedule/plan for dealing with that data set
4. perform the work scheduled for the first hour of your schedule
5. reflect on the experience.

Point Value: In total, this assignment is worth 50 points (out of 500) towards your final grade.

General Rules on The Assignment

1. **Format: Work** to create a **coherent report** that fulfills the requirements below. **Do not just check off the items with chunks of code.**
2. **Due Date:** This assignment should be turned in by the *end of the day* (i.e. midnight) on **Wednesday, November 18**. If you realize that you will not meet this deadline, you must e-mail Dr. Penland with a proposed extension utilizing the proposed number of your 5 "day-passes" for the semester. **If neither a submission nor an extension request is received by Wednesday, November 18, you will receive a "0" for the assignment.**

   (a) **General Statement on Academic Integrity**
   While individual competence is important, so is the ability to gather information and work as part of a team. The rules below are intended to represent this reality. **They are not "anything goes"** – notice that there are still clearly defined instances where Academic Dishonesty can occur. **If you are uncertain whether or not something is allowed, it is better to e-mail Dr. Penland and ask first.**

      i. **Rules on Collaboration**: You may work with up **two** co-authors on this project. Teams of co-authors will be **Please respond to the appropriate assignment with the name of your co-author by no later than the end of the day on Monday, November 16**. All people in the group must indicate co-authorship status for it to be accepted.

      - **Example #1:** Alice and Danielle want to work together. Each one responds to the Coauthor Activity indicating this. They will be placed in a single group, submit a single file, and receive the same grade.
      - **Example #2:** Bob really wants to work with Charlie, but Charlie would prefer to work by himself. Bob responds to the activity and indicates Charlie is his co-author. Charlie indicates he is working by himself. Bob and Charlie will not be co-authors; each one will work by himself.

- **Example #3:** Ryan, Samantha, and Timothy are good friends who often do projects together. Ryan lists Samantha as his co-author, Samantha lists Timothy as her co-author, and Timothy lists Ryan as his co-author. **This is invalid; each one will end up working by themself.**
- **Example #4:** Marilyn, James, and Elvis have a strange social dynamic. James indicates that he is working with Marilyn and Elvis. Marilyn and Elvis each indicate that they are working with each other. James will be a group of one; Marilyn and Elvis will be a group of two.

These examples should give you the idea to generalize how other situations will proceed. **If uncertain, please e-mail Dr. Penland.**

ii. **Rules On Communication:** You may discuss this project with anyone you wish, whether or not they are a co-author and whether or not they are in this class. Any such discussions must be documented in the notebook. You **may not** share code with people who are not listed as co-authors. **Sharing code in this way is Academic Dishonesty, whether or not the sharing is acknowledged.**
You **may** always communicate with Dr. Penland, but these should be acknowledged as well.

- **Example #1:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie sends him a link to a relevant section of the `matplotlib` documentation. This is acceptable, **as long as Bob acknowledges Charlie's help in the appropriate section on the Jupyter notebook.**
- **Example #2:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie refers Bob to the portion in the Blackboard lecture videos where this was discussed. This is acceptable, **as long as Bob acknowledges Charlie's help in the appropriate section on the Jupyter notebook.**
- **Example #3:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie writes an example, screenshots it, and sends it to Bob. **This is unacceptable and constitutes Academic Dishonesty.**
- **Example #4:** Bob is stuck on how to change the color in a scatterplot using `matplotlib`. He asks Charlie. Charlie sends a Jupyter notebook to Bob that includes this part of code. **This is unacceptable and constitutes Academic Dishonesty, even if Bob says "I got the code for how to change color in a scatterplot from Charlie".**

iii. **Rules On References:** You may utilize any coding resources you wish, except for code that has been created by other students in the class who are not your co-authors. **Any reference** from which you use information must be cited using the appropriate format. **Any code that is utilized from a reference must be cleared for the appropriate license, and acknowledgment must be given in the notebook close to the place where the code is listed.** You will probably need to modify code examples that you reference; the reference should still be included.

Rubric: The points will be assigned based on the following criteria.

- **(0 points total on the assignment; all other items irrelevant.)** on the assignment for any Academic Dishonesty violations as outlined above. **This includes no information given for acknowledgments or resources.**
- **(BONUS: +3 points)** if you complete the Project with a co-author whom you did not know before November 12, 2020.
- **(BONUS: +3 points)** if you ask a question that helps Dr. Penland improve the clarity of the instructions, or improve future projects in some way.
- An Acknowledgments section is completed that refers to any communications with anyone who was not a co-author (including Dr. Penland). **Part #1:** Please answer these questions in a file called "XX-P3-background.pdf". How you create the .pdf file (using Jupyter notebook, Microsoft Word, LaTeX, etc.) is up to you. **Please notice that this part should contain no code.**
- **(+1 points)** Choose a data set that you have not used on either Project #1 or Project #2. It is acceptable to use data that you have previously seen in a homework problem.
- **(+1 points)** Clearly identify and describe the features, including whether they are categorical or numerical. If there are a very large number of features (more than 20), you may "chunk" them into types. (For instance, in the `spam` e-mail data set, there were a large number of "word_freq" variables which could be described in a single bullet point.)
- **(+1 points)** Clearly identify and describe the target.
- **(+1 points)** Describe any processing that needs to be done on the data to make it work for your Machine Learning algorithm. How are you going to encode the data?
- **(+1 points)** Explain whether this is a classification or a regression problem.
- **(+1 points)** Explain clearly the *metric* by which you will evaluate performance on this problem. This may include **satisfiability metrics** as well as **optimization metrics.**
- **(+1 points)** Give an estimate of the *Bayes Optimum* on this problem for this metric. Justify your estimate, using anything you know about the problem, machine learning, etc. **Part #2:** Please perform this part in a file "XX-P3-planning.pdf". How you create the .pdf file (using Jupyter notebook, Microsoft Word, LaTeX, etc.) is up to you. **Please notice that this part should contain no code.**
- **(+10 points)** Suppose that you were given one eight-hour day to build a working model for this data. Give a detailed schedule for how you would spend this time. This schedule could be detailed enough so that if you needed to call in sick on the "model-building" day, the rest of your team could follow your plan. For instance, here are some questions to ask yourself:

  - what portion of the day would you spend on trying to determine useful features by hand? How would you do this?
  - How much time would you spend plotting the data? Which plots would you make? What are you hoping to find out through these plots?
  - How long would you give yourself to train various models? Which types of models would you try to train? How would you do this?
  - How long would you give yourself to compare different models? Which techniques will you use (residuals, confusion matrix, etc.)?

The final plan should be a schedule, along with a detailed description of the tasks to perform for each item. **Part #3:** Please perform this part in a file called "XX-P3-work.ipynb".

- **(+10 points)** Set a timer and work the first hour of your eight-hour plan. Record your work in a Jupyter notebook. (It does not need to be well-organized.) **Part #4:** Please perform this part in a file called "XX-P3-reflection.pdf". How you create the .pdf file (using Jupyter notebook, Microsoft Word, LaTeX, etc.) is up to you. **Please notice that this part should contain no code.**
- **(+4 points)** How accurate was your time assessment? Was there anything that went significantly longer, or significantly shorter, than you expected?
- **(+3 points)** Look at the Jupyter notebook that you compiled during the first hour of work. If you were forced to leave the project now and give this plan and this notebook to another person on your team, how useful do you think it would be? What information is missing?
- **(+3 points)** List three specific things that you learned from this assignment.
- **(+10 points)** This assignment was noticeable different than the previous projects. Do you think that this project would be useful to include in future Math 475/575 courses? If yes, why? If no, why not? Are there any specific changes that you would make to this project to make it more informative? x

---