Spring 2022 Project Specifications

You will work in groups of 4. The group project will be to use the methods learned in class to perform classification on a dataset of your own choosing. You will try a variety of classification algorithms on the data and compare their results. You will want to use the appropriate metrics to measure each classifier's performance and you will want to use techniques learned in class to try to improve those metrics.

You will not explicitly be graded on the performance of your classifiers. In other words, if you are only able to get a 60% accuracy on your dataset, that is ok - a low accuracy (or F-score) will not affect your grade.

You will be graded on your methodology: What is the process you used, and what things did you do to try to improve the performance?

Final Deliverable:

You will submit a Jupyter Notebook that walks a reader through your analysis. The notebook should effectively communicate the "story" of your data analysis: not just what you found, but *how* you found it.

You should start with a brief introduction to the problem (utilize markdown cells for this). What is the data science problem you are trying to solve? Why does the problem matter? What could the results of your predictive model be used for? Why would we want to be able to predict the thing you're trying to predict? Then describe the dataset that you will use to tackle this problem.

From here, proceed with the data science. Perform data cleaning, data exploration, feature engineering, etc. Use the markdown cells to describe and clarify each part of the process, so that a reader can easily follow along with what you've done.

Then move on to the modeling, and continue to use the markdown cells to walk a reader through your process.

Finally, display and analyze your results. Include a final conclusion - this may be an analysis of which model worked best, or which feature engineering worked best, or it may be any interesting insights you discovered about your data, or anything at all that you want to conclude from your work.

Project Grading Rubric:

Your grade will be determined based on the following rubric:

10 points	**Data Prep:** Was data cleaning performed thoroughly and correctly?
10 points	**Data Exploration:** Was data sufficiently explored?
20 points	**Feature Engineering:** Were features engineered appropriately and correctly?
40 points	**Data Analysis**: Were correct data analysis and machine learning techniques used? Is the code correct?
10 points	**Code & Notebook Quality:** What is the quality of the code? What is the quality of the notebook? Is the notebook well narrated? Are appropriate outputs and/or markdown cells used to communicate the process and results?
10 points	**Outcome:** Was the prediction problem set up appropriately for the data? Was the analysis complete? How well was the data analyzed to solve the problem? Are the results interpreted correctly?
TBD (bonus points)	**Difficulty:** How difficult was the dataset to work with? Additional points may be added at the instructor's discretion, for datasets that are more difficult to work with.
TBD (point deduction)	**Peer Evaluation:** Each team member will assess the contribution of the other members of their team. Peer assessments of your contributions will be used to adjust your final grade at the instructor's discretion. Any group members who do not participate in the project at all will receive a 0 for their project grade.

Where to find data?

For this project, you may select any dataset of your choosing, **except** datasets found in the UCI Machine Learning Repository.

There is SO MUCH data available! This can actually make it quite difficult and daunting to find a "good" dataset to analyze.

Here are a few datasets that I particularly like - you are welcome to choose one of these:

- Austin Animal Center data: predict which animals will be adopted [**Intakes (https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Intakes/wter-evkm)** and **Outcomes (https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238)**] (tables will need to be merged on ID)

- Speed Dating data: predict who will be asked for a second date [**dataset (https://www.kaggle.com/annavictoria/speed-dating-experiment)**]

- Covid-19 data by county: These datasets contain the number of covid cases and number of covid deaths by county. There are also datasets on mask usage by county. These could be combined with all kinds of other data available by county (things like population demographics, poverty rates, education levels, etc.) to predict covid cases, or covid deaths, or mask usage rates, etc. Because we did not cover regression in this class (predicting a continuous number), if you'd like to predict something like covid deaths, or covid cases, etc, those numbers could be binned to make it a discrete classification project. [**covid-19 datasets** **(https://github.com/nytimes/covid-19-data)** ; various **other datasets by county** **(https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/)** , including population, poverty rates, education rates, unemployment rates; you can find many other datasets of things by county!]

Or you may choose to find your own dataset! Here are some places that you can look:

KD Nuggets: List of Data Sets **(https://www.kdnuggets.com/datasets/index.html)**

Data.gov **(https://www.data.gov/)**

Enigma **(https://public.enigma.com/)**

Police Data Initiative **(https://www.policedatainitiative.org/datasets/)**

Harvard Dataverse **(https://dataverse.harvard.edu/)**

20 Free Sports Datasets for ML **(https://lionbridge.ai/datasets/20-free-sports-datasets-for-machine-learning/)**

Forbes: 33 Brilliant and Free Data Sources Anyone Can Use **(https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#6c41a798b54d)**

====================

Additional Resources:

The Comprehensive Guide for Feature Engineering **(https://adataanalyst.com/machine-learning/comprehensive-guide-feature-engineering/)**

Feature engineering: **Part 1 - Continuous Data** **(https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b)** , **Part 2 - Categorical Data** **(https://towardsdatascience.com/understanding-feature-engineering-part-2-categorical-data-f54324193e63)**

How to Handle Class Imbalances **(https://elitedatascience.com/imbalanced-classes)**

Essential Checklist for Any Data Analysis Project **(https://blog.k2datascience.com/essential-checklist-for-any-data-analysis-or-science-project-7c4fa924e563)**

Nine rules *not* to follow **(http://www.kdnuggets.com/news/2004/n08/18i.html)** : Common pitfalls when first attempting to tackle data mining problems.

An Introduction to Machine Learning with scikit-learn (https://scikit-learn.org/stable/tutorial/basic/tutorial.html) : This tutorial has multiple parts - click through it with the 'next' button in the bottom right.

Machine Learning Performance Improvement Cheat Sheet (https://machinelearningmastery.com/machine-learning-performance-improvement-cheat-sheet/)

Hyperparameter Tuning for Machine Learning Models (https://www.jeremyjordan.me/hyperparameter-tuning/)