Final Project Guidelines

For the final project, you will perform an open-ended machine learning analysis of a healthcare dataset of your choice. You can work alone or in groups up to three people. Please send us your teams by **11/9/18 @ 11:59 PM**. This assignment will be 40% of your total grade. The final project is due by **12/7/18 @ 11:59 PM** but we highly recommend you start working on it early.

Dataset Selection & Planning

We have provided a list of sites to search for datasets. Your group will select a healthcare dataset in an area you are interested. Send us a short paragraph identifying your dataset and list out the potential models you would like to train. Identify any potential issues and challenges you may run into. Datasets should have at least 30 samples in each class, and at least 10 features.

Dataset selection is due **11/20 by 11:59 PM.**

Note: The due date for selecting a dataset is 2 weeks before the due date of the final project. Start ASAP!

Some helpful dataset sites to search:

- https://www.kaggle.com/tags/healthcare
- https://www.data.gov/health/
- https://healthdata.gov/
- https://toolbox.google.com/datasetsearch
- *Additional Links will be Added*

Jupyter Notebook Analysis

The structure of your analysis will be similar to the past two codelabs you have done. We highly suggest you to use the past two codelabs as reference. We will be hosting office hours on Mondays and Wednesdays or by appointment. Please let us know if you are having issues.

There are 4 separate steps listed below. Each section will be worth 20% of the project grade.

Data Preprocessing	20% ● Remove all null values and make sure all data fields are valid ● convert categorical variable to indicator variables ● understands labels for fields and what results represent
Data Visualization/Exploration	20% ● Identify the feature breakdown of your dataset (number of malign/benign patients, male/females/age range etc.) ● Find linear correlations between features (You can use the charting libraries we provided in previous codelabs or ask us) ● At least 5 graphs including a covariance correlation.
Feature Selection	20% ● Select features to train/tune based on previous step ● Explain why you chose these features to tune (you do not need to remove these right away, may impact accuracies)
Machine Learning	20% ● Select at least three binary classifier models to tune (logistic/linear regression, random forest, SVM, decision trees, naive bayes, etc). ● Tune parameters and features (You can start off with training on all features) ● Test resulting accuracies with cross validation ● Explain your results, which models performed the best with tuning?

	(identify the bias/variance tradeoff for each model and select the best one at the end of tuning).

Presentation

Your team will present your finding to the entire class on the last day of lecture. You will describe your dataset, identify the features contained within the dataset, show any interesting correlations you found from data visualization/feature selection, and explain the steps you took to tune your models and the results you achieved. This will be more open ended so feel free to cover additional topics. Presentations will be approximately 5 minutes per team. Submit your presentation with your final project.

Grading

Dataset Selection & Plan	10%
Jupyter Notebook Analysis	80%
Presentation	10%