SML 312 **Mini-Project #2 -- Classification** **Spring 2022**

The following mini-project is focused on sharpening your Data Science skills in the context of Classification! Your writeup should be submitted as a well-documented Python / Jupyter Notebook by e-mail to jonhanke@princeton.edu and sulinl@princeton.edu by 10pm (EST) on Wednesday March 16th, 2022.

Collaboration: All class projects are required to be solely your work, with no collaboration on modelling or design choices, and you will be required to certify this when you submit your assignment. If you want to ask others about or share general references, please feel free to post them on Slack where all comments are clearly visible.

References: Please feel free to use online resources (with clear inline citations in your code) to better understand the process of using your basic data science tools (Python / Jupyter / Pandas / Scikit-Learn) to perform a specific task (i.e. make a scatter plot, fill in missing values, etc.), but ***you may not consult*** online references specific to your given dataset, or solicit/receive comments outside of course discussions about the choices needed to perform EDA or modelling specific to your dataset. When in doubt about a reference, please feel free to ask on our Slack channel, or during Class / Precept / Office Hours.

Mini-Project Questions:

1. Comparing Classification Models (70%)
    a. Please use the scikit-learn breast cancer dataset (e.g. with the Python `sklearn.datasets.load_breast_cancer()` command), and perform a randomized (20%/80%) test/train split of the data to allow for model cross-validation in later steps.

    b. Please perform an exploratory data analysis on your training set to gain intuition for your modelling efforts. Do you notice anything interesting? Which features seem most important? Why?

    c. Please create a binary classification model on your training set to predict which data samples are benign or malignant, and make a confusion matrix to report your results.

    d. For this classification problem, what do you feel is the most appropriate way to measure the "goodness" of the model? Please carefully explain your reasoning!

    e. Please repeat part 1c for each of the following types of classification models, choosing parameters that seem appropriate to give a good model. For each

model you make, record the confusion matrix and how well the model performs for the “goodness” metric you decided on in part 1d:

i. Nearest Neighbors
ii. Naive Bayes
iii. Logistic Regression
iv. Support Vector Machines
v. Decision Trees
vi. Random Forests

f. Which model performed the best on the training data, and how well did they perform in general?

g. Which model performed the best on the testing data, and how well did they perform in general?

h. Compare the model performance on the testing and training data to see if any of your models were overfit!

2. Explaining your Model (20%)

   a. Please explain which features are the most important in each of your models in part 1e? Is this similar across all models? How does this agree with your expectations from part 1b? Please carefully explain your reasoning!

   b. Of all of the models in 1e, which was the most explainable? Why?

   c. How might this information be useful in communicating to doctors performing screenings for cancer based on these images?

3. Varying the Decision Threshold (10%)

   a. For one of the models above that predicts the probabilities to perform the classification (i.e. Naive Bayes, Logistic Regression, Decision Trees, Random Forests), please construct/plot the Receiver Operating Characteristic (ROC) curve to show how well the model performs with across various choices of thresholds.

   b. What is the area under this ROC curve? What does this tell us about our model?

   c. **Extra Credit:** What is the optimal choice of threshold for this model to optimize your “goodness” measure in part 1d? How did you find this?