Cogs 118A Final Project Proposal

- Goal
    - We want to perform binary classification tasks by using 6 algorithms on 6 datasets, run 5 trails for each algorithm by hyperparameter tuning, and use 7 metrics to evaluate and compare performance of those algorithms

- 6 datasets to be used, references for where they are obtained, description of dataset, and whether it needs cleaning or not
  We would need to clean all datasets
    - Steam gaming dataset - contains information about games found on digital game distributor Steam, such as the number of owners, genres, categories, prices, and many other features of the game. We will classify whether the game is popular or not based on a threshold.
      https://data.world/craigkelly/steam-game-data
    - Mushroom classification - contains data about different mushrooms and their properties. We will predict whether the mushroom is edible or not based on these properties. https://www.kaggle.com/uciml/mushroom-classification
    - Movie rating - Using the genres of a movie to predict the maturity rating of the movie (rated R, PG-13, PG, etc). Make columns IsRatedR, IsDocumentary, IsShort, etc. Predict the maturity rating of the Movies as IsRatedR or not.
      https://www.kaggle.com/samruddhim/imdb-movies-analysis
    - New York City Airbnb Open Data - includes information about hosts, geographic location, neighborhood, room type, and the price. We would choose a threshold of price and predict whether the price is above the threshold and the house can be classified as expensive or not.
      https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data?select=AB_NYC_2019.csv
    - Rain in Australia - contains about 10 years of daily weather observations from many locations across Australia. We want to use those weather observations to predict whether it would rain in the next day or not.
      https://www.kaggle.com/jsphyg/weather-dataset-rattle-package/discussion
    - Dataset on Cardiovascular Disease presence - contains age, gender, and other features that may contribute to cardiovascular disease dataset. We want to predict whether a person has cardiovascular disease or not.
      https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

- 6 algorithms being used, and in what library/package they are implemented
    - Decision Tree
      from sklearn.tree import DecisionTreeClassifier
    - Logistic Regression
      from sklearn.linear_model import LogisticRegression
    - Perceptron Classifier
      from sklearn.linear_model import Perceptron
    - Support Vector Machine

from sklearn import svm
    - K Nearest Neighbors
      from sklearn.neighbors import KNeighborsClassifier
    - Random Forest Classifier
      from sklearn.ensemble import RandomForestClassifier

- 7 metrics being used
    - Accuracy
    - Precision
    - Sensitivity/Recall
    - Specificity
    - F1 score
    - ROC curve
    - Cross Entropy

- list of group members and what aspects of the project each group member will be primarily responsible for
    - Duy Pham (A15467782)
        - Clean Steam dataset and the mushroom dataset
        - Run the 6 algorithms and calculate the 7 error metrics on each dataset
    - Yunchun Pan (A15195894)
        - Clean Rain in Australia dataset and Cardiovascular dataset
        - For Rain in Australia dataset, I would use 6 classification algorithms to predict whether it would rain in the next day using weather observations of a location, run 5 trails for each algorithm by hyperparameter tuning, and use 7 metrics to evaluate and compare performance of those algorithms
        - For Cardiovascular dataset, I would use age, gender, and other factors to predict whether the person has this type of disease, run 5 trails for each algorithm by hyperparameter tuning, and use 7 metrics to evaluate and compare performance of those algorithms
        - I would share my progress with other group members during weekly meeting, improve my codes, and write the final report
    - Melchisedec Lee (A15597085):
        - Clean 2 datasets “Movie Rating” and “NYC AirBnB” before next meeting
        - Perform the 6 algorithms and 7 error metrics on respective datasets
        - 5 trials for each algorithm with hyperparameter tuning