**CIS 4526: Foundations of Machine Learning**
**Homework 2: Decision Trees**

**Instructor: Alex Pang**
**Due Date: 4/7 23:59 pm**

# Overview

The goal of this homework is to implement Deccision Tree and Random Forest yourself. I will give you a script (*decision_tree_orig.py*) that follows a blog, namely,Implementating a Decision Tree from Scratch, where the author explains how you can implement a Decision Tree from scratch. However, it has deficiency and your job is to make it better. Download the python file decision_tree.py and a jupyter notebook for testing from Canvas and fill in the missing parts. Your job is to fill in the missing code in decision_tree.py.

# What you need to do

Fill in the missing code inside the TODO sections. In particular, add the following parts

1. Add a _gini method to calculate gini index so the model can switch between using entropy and gini as the criterion.
2. Make the code work for the case where the target variable can be a categorical variable as well as integer variable from encoding the different classification classes.
3. Add classification_report and confusion_matrix functions (do not call the sklearn versions).
4. Implement the RandomForest model (do not call the sklearn version).
5. Add an additional stopping criteria based on whether the impurity (entropy or gini) is low enough in addition to max_depth and min_samples_split. The idea is have a parameter to control overfitting.
6. Add reasonable enough comments to each methods.

Zip up the following files listed below and submit the zip file to Canvas.

1. breast_cancer.csv under the data subfolder
2. decision_tree_orig.py
3. your modified version of decision_tree.py
4. ML_SP24_HW2.ipynb with your answer inside the notebook

# Grading criteria

1. Correctness
2. Good enough comments
3. Elegance
4. Oral explanation (possible)