Homework 4 **CSCE 633**
Due: 11.59pm on November 4, 2020

Instructions for homework submission
Please submit on eCampus a **single pdf** file containing your solutions.
a) Please write a brief report and **include your code right after each answer**.
b) For each answer, please explain your thought process, results, and observations. Please do not just include your code without justification.
c) Create a **single pdf** and submit it on **eCampus**. Please do not submit .zip files or colab notebooks.
d) The maximum grade for this homework is **6 points** (out of 100 total for the class).

Question: Decision Tree and Random Forest

Classifying benign vs malignant tumors: We would like to classify if a tumor is benign or malign based on its attributes. We use data from the Breast Cancer Wisconsin Data Set of the UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`.
Inside "Homework 4" folder on Piazza you can find one file containing the train data ("hw4_train.csv") and test data ("hw4_ test.csv") for our experiments. The rows of these files refer to the data samples, while the columns denote the features (columns 1-9) and the outcome variable (column 10), as described bellow:

1. Clump Thickness: discrete values $\{1, 10\}$
2. Uniformity of Cell Size: discrete values $\{1, 10\}$
3. Uniformity of Cell Shape: discrete values $\{1, 10\}$
4. Marginal Adhesion: discrete values $\{1, 10\}$
5. Single Epithelial Cell Size: discrete values $\{1, 10\}$
6. Bare Nuclei: discrete values $\{1, 10\}$
7. Bland Chromatin: discrete values $\{1, 10\}$
8. Normal Nucleoli: discrete values $\{1, 10\}$
9. Mitoses: discrete values $\{1, 10\}$
10. Class: 2 for benign, 4 for malignant (this is the **outcome** variable)

(1) (1 point) Data exploration: Using the training data, plot the histograms of the class outcome and each feature (i.e., 10 histograms total). Compute the number of samples belonging to the benign and the number of samples belonging to the malignant case.

(2) (1 point) Conditional entropy: *Implement* a function that computes the conditional entropy of each feature, conditioned on the class outcome. Using the training data, compute the conditional entropies for each feature (i.e., 9 values total). Which features are the most discriminative of the outcome?

Hint: For implementing the conditional entropy, please follow the example that we discussed in class.

(3) (2 points) Decision tree classification: Use a decision tree classifier to classify between benign and malignant tumor based on the features provided. Identify the optimal hyper-parameters (e.g., tree depth) using hyper-parameter tuning through a 5-fold cross-validation on the training set. Report the classification accuracy for all hyper-parameters from the cross-validation process on the training set, as well as the classification accuracy on the test set using the best hyper-parameter from the cross-validation.
Note: You can use any available library for the decision tree and the cross-validation.

(4) (2 points) Random forest tree classification: Repeat the same task as in question (3) using a random forest classifier. Experiment with the optimal tree depth and number of trees. Compare and contrast the performance of the decision tree with the random forest classifier.
Note: You can use any available library for the random forest and the cross-validation.