CSCI 184 Homework 3
Due Date: May 18th, 2022

1) [60 points] Predicting Purchases from Social Network Ads

What do you need to do?

1. Load the dataset from 'Social_Network_Ads.csv' into a pandas DataFrame and print it along with its shape. 'Purchased' is the target variable.
2. Print the column names and the data type of each column.
3. Normalize the features in the input dataset.
4. Perform Random Forest Classification using sklearn. Make sure to enable bootstrap and oob_score
5. Find the **oob error** for different numbers of trees. Figure out which parameter sets the number of trees.
6. Plot the **oob error vs number of trees.** (Similar to example in class) and find the number of trees with the lowest error. If multiple, report all.
7. Add the plot from step 6 in your report and also report the number of tree(s) with the lowest error.
8. Submit your code as an .ipynb file and a document reporting your findings.

2) [40 points] Theory

Please answer the following in your own words.

1. Joey was using the decision tree model to classify credit card transaction records into legitimate and fraudulent. He collected a dataset with 10000 records, which consists of 5000 legitimate and 5000 fraudulent. He randomly selected 8000 records to learn a decision tree model and evaluated the decision tree model on the remaining 2000 records. In evaluation results, he found out that the training error rate was 1% but the test error rate was 20%.
    a) [5 points] Why did this happen? State all the reasons you can think of discussed in class. You can also add reasons you find elsewhere.
    b) [5 points] What would you suggest Joey do next to solve this issue? State all the ways discussed in class. You can also add suggestions you find elsewhere.

2. [10 points] Explain what is Bagging? What is Bootstrapping? Also draw a figure to explain the process.
3. [5 points] Explain the working of the Random Forest Algorithm with the help of a diagram.
4. [5 points] Why do we prefer a forest (collection of trees) rather than a single tree?
5. [5 points] How do we determine the ‘correct’ number of trees in Random Forest? Explain with an example.
6. [5 points] We are using both validation set and test set to find the performance of the model so then isn't the validation set the same as the test set? Justify your answer.