CS 4661: Introduction to Data Science
Homework2
Due Date: Fri, Oct 11

Question1: Decision Tree for Weather Forecasting:
Suppose that we want to build a *decision tree classifier* to perform weather forecasting! We have 3 features (Temp, Humidity, Wind), and a binary Label (sunny/rainy). The following table includes the data collected over the past two weeks.
Based on this data, which feature is the best feature to put on the top of the tree? Justify your answer by providing detailed Entropy calculations.

Temp	Humidity	Windy	Label
high	low	Yes	Sunny
low	high	Yes	Rainy
high	low	No	Sunny
high	high	No	Sunny
mild	mild	No	Sunny
mild	high	No	Rainy
low	mild	Yes	Rainy
low	low	Yes	Rainy
low	high	Yes	Rainy
high	low	No	Sunny
high	mild	No	Sunny
mild	mild	No	Sunny
mild	high	No	Rainy
low	mild	Yes	Rainy

Question2: KNN Classification in sklearn
Write and submit your python codes in "Jupyter Notebook" to perform the following tasks (submit the .ipynb file). Make sure to provide proper descriptions as MarkDown for each section of your code (each section of the code must have a short meaningful description right before that, describing what this part of the code is supposed to do!)

a- Read the iris dataset from the following URL:
https://raw.githubusercontent.com/mpourhoma/CS4661/master/iris.csv
and assign it to a Pandas DataFrame as you learned in tutorial Lab2-3.

b- Split the dataset into testing and training sets with the following parameters: **test_size=0.4, random_state=10**

c- Instantiate a KNN object with K=3, train it on the training set and test it on the testing set. Then, calculate the accuracy of your prediction as you learned in Lab3.

d- Repeat part (c) for K=1, K=5, K=7, K=15, K=27, K=59 (you can simply use a "for loop," and save the final accuracy results in a list). Does the accuracy always get better by increasing the value K?

e- Now, suppose that we would like to make prediction based on only **ONE single feature**. To find the best single feature, we have to try every feature individually. In other word, we have to repeat part (c) with K=11, four times (each time using only one of the 4 features), and compute the accuracy each time. Then, check the accuracies.
Which individual feature provide the best accuracy (the best feature)? Which one is the second best feature? (**Note**: you have to train, test, and evaluate your model 4 times, each time on a dataset including only one of the features, and save the final accuracy results in a list).

f- Now, we want to repeat part (e) (with K=11), this time using **two features**. you have to train, test, and evaluate your model for 6 different cases: using (1^{st} and 2^{nd} features), (1^{st} and 3^{rd} features), (1^{st} and 4^{th} features), (2^{nd} and 3^{rd} features), (2^{nd} and 4^{th} features), (3^{rd} and 4^{th} features)!
Which "**feature pair**" provides the best accuracy?

g- **Big Question:** Does the "best feature pair" from part (f) contain of both "first best feature" and "second best feature" from part (e)? In other word, can we conclude that the "*best two features*" for classification are the *first best feature* along with the *second best feature* together?

h- Optional Question: Justify your answer for part (g)! If yes, why? If no, why not?