Assignment 4:
Due Date: 11.59 PM 12/02/2022

1 Question 1 (30 pts)

: In this question we are going to design a machine learning problem for credit card fraud detection. You have to predict, based on certain attributes, whether a transaction is fraudulent or legitimate. Answer the following questions.

- 1 a. Is this a classification problem or a regression problem? (2 pts)
- 1 b. Describe at least 4 discrete features of each transaction that is going to help in our problem. (8 pts)
- 1 c. Describe at least 3 continuous features of each transaction that is going to help in our problem. (6 pts)
- 1 d. If we were to use an ML algorithm that only takes discrete features as input, what can we do to the continuous features you described in 1 (c) so that they can be used by the ML algorithm? (4 pts)
- 1 e. How would you know if your model suffers from overfitting? If you are using decision tree in your problem, how would you deal with overfitting? (4 pts)
- 1 f. In Table 1, observe the data and use your own intuition to draw a decision tree that can classify the data into Class 1 and Class 2. (6 pts)

2 Question 2 (40 pts)

:

1. The input dataset is : **ballons.csv**
2. The output file is: **output_ballons.txt**
3. Starter code: **q2_decision_tree.py**

You can find the dataset and starter code in **CMSC_471_Assignment_4.zip** on blackboard.

You have to implement the code to choose the best feature to split the dataset in a decision tree. Please use entropy and information gain as the metrics to choose a feature. You will find some starter code in **q2_decision_tree.py**. Implement the functions that are marked with '#TODO'.

Note: Your python code should **NOT** take any input parameter. It should print the output in a file 'output.txt'. If your python code fails to run without any input parameters, you will **NOT** be graded for this question.

Table 1: Table with variables A,B,C and Output Label

A	B	C	Output
1	1	1	Class 1
1	1	0	Class 1
0	0	1	Class 2
1	0	0	Class 2

3 Question 3 (30 pts)

1. The input training dataset is : **titanic_train.csv**
2. The test dataset is: **titanic_test.csv**
3. The output file is: **output_titanic.txt**

For this question, consider the Titanic dataset. You have to predict whether a person will survive based on the features of that person.

You may use sklearn's models for training. Please use **titanic_train.csv**. for training and report your scores on **titanic_test.csv**.

If you want to reduce the complexity, you may drop the columns 'Name', 'Ticket', and 'Cabin'. Please note that the feature 'Passenger Id' is an identifier field. For fields with missing values, use np.mean or np.median to impute the missing values. For example if the field age has 5 values 10,10,20,20, [missing], use 15 to impute the missing value.

Train the following ML models (at least) on the training dataset

- Random Forest (10 pts)
- Logistic Regression (10 pts)
- SVM (10 pts)

For each of these models, submit your classification report (use sklearn.metrics.classification report) on the test dataset (test.csv). Write your classification report to the **'output_titanic.txt'** file.

You can find the datasets in **CMSC_471_Assignment_4.zip** on blackboard.

Note: Your python code should **NOT** take any input parameter. It should print the output in a file **'output_titanic.txt'**. If your python code fails to run without any input parameters, you will **NOT** be graded for this question. You have to ensure that the file that you read (train or test) is in the same directory as your code.

Extra Credits (20 pts): Explore other models in the sklearn library, run them and report any other model that performs better than the best performing model that you encountered in Question 3. Report the classification report on the test dataset for this model.