ITP 449, FALL 2020

HOMEWORK 6
20 POINTS

For each one of the following questions, write Python code in PyCharm.

- For each question, create a *new* Python file. Name each *lastname_firstname_hw#_q#.py* etc.
- Create a header in each file using *comments* to display your name and HW information. After that write your Python code.
  # Tommy Trojan
  # ITP 449 Fall 2020
  # HW6
  # Q1

- Apart from the above comments, include single line comments describing the core logic of your algorithm / code.
  As an example,

  #Creating a DataFrame using the csv file.

The dataset[1] you will analyze in this HW (available on the blackboard) is the RMS Titanic. https://en.wikipedia.org/wiki/RMS_Titanic. Your goal is to classify survivability based on the various factors of the passengers. These factors are listed below:

Variable	Definition	Key
survived	Survival	No, Yes
class	Ticket class	1st, 2nd, 3rd
sex	Sex	
Age	Age	Child, Adult

[1] https://www.kaggle.com/c/titanic/data

Problem #1

1. Read the dataset into a dataframe. (1)
2. Explore the dataset and determine what is the target variable. (2)
3. Drop factor(s) that are not likely to be relevant for *logistic* regression. (2)
4. Make sure there are no missing values. (2)
5. Plot *count* plots of each of the remaining factors. (2)
6. Convert all categorical variables into *dummy* variables. (2)
7. *Partition* the data into train and test sets (70/30). Use *random_state = 2020*. (2)
8. *Fit* the training data to a *logistic regression* model. (2)
9. Display the *accuracy* of your predictions for survivability. (2)
10. Display the *confusion matrix* along with the labels (Yes, No).

    Hint: You may want to use from *sklearn.metrics import plot_confusion_matrix* (2)
11. Now, display the predicted value of the survivability of a *male adult passenger traveling in 3rd class*. (3)