CS 250 01 Spring 2025

Extra credit class assignment to be done on 3/28/25

Due on Moodle: By midnight on 3/28/25. There is no time limit for the assignment, but it will remain open for 24 hours. There will be no late submissions.

For this class assignment, download the Titanic dataset (link) and do the following:

1.
    a. Read the data into a dataframe with the PassengerId column as index_col. [2 points]
    b. Separate the **'Embarked'** column and convert that into an array. This will be your dependent variable (NOT the survived column). [2 points]
    c. Drop 'Embarked','Ticket', 'Cabin', and 'Name' columns from the main dataframe. The remaining columns will form your independent variable. [3 points]
    d. Convert the 'Sex' column to a numerical column (0/1). (code given in class) [3 points]

2.
    a. Split the X and y data into 70% training and 30% testing using train_test_split() [2 points]
    b. Train a Random Forest classifier on the data. [2 points]
    c. Print the accuracy score of the classifier. [1 points]
    d. Evaluate the performance on the test data using a normalized confusion matrix (code given in class). The matrix should have the class labels "S", "C" and "Q" on the axes. [5 points]