http://localhost:8888/?token=00ed496f0fc033ea5616d20f4b1e888e2ea7f12046b50522&token=00ed496f0fc033ea5616d20f4b1e888e2ea7f12046b50522CMP 464 & 788 Mid-Term Project – Stage 1

For the mid-term project, we will tackle the Titanic dataset on Kaggle.com. This project will be completed in three stages: 1) Data preparation, 2) Model construction, and 3) Model find-tuning and analysis. **The first stage is due on Monday, March 12 at 11:59pm.** You are expected to submit a complete Jupyter notebook with python code (with detailed comments), numerical results, and graphs (with proper labeling) to Blackboard.

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although where was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this project, you are asked to apply machine learning tools to analyze the passenger information from the training set, and try to predict which passengers in the test set survived the tragedy.

1. Download the training data and test data from https://www.kaggle.com/c/titanic/data as .csv files

2. Load the two datasets as pandas DataFrames. Combine them into a single DataFrame. For the purpose of this stage, we will only use the combined dataset.

3. Obtain basic information of variables. For each variable, find out:
1) What does this variable represent
2) The meaning of values
3) Numerical summary
4) Graphical distributions of values

4. Data Cleansing
Are there missing values in each column? If so, is it possible to impute the missing data in a reasonable way? Discard all passengers with incomplete information may significantly reduce the number of data example. We should use different methods to handle the missing data according to the nature of the variable:
1) If a variable only has a tiny fraction of values, a common practice is imputing with mean or median value among passengers belonging to the same group.
2) If most values for a variable are missing, it is usually better to simply discard this feature.
Please go through all variables and handle the missing values.

Are there outliers? If so, decide whether these values should be removed.

5. (optional) You may think that the name feature is irrelevant. However, a closer look at the names will show that some passengers have rare titles such as Dr. Mme, or Master., which indicates their social status. Please create a column to indicate whether a passenger has a rare title.