*Session 17: Assignment 17.1*

Problem Statement : - Predicting Survival in the Titanic Data Set

Step 1 - Define problem The goal of this tutorial is to predict if a passenger on the Titanic will survive based on a number of variables. For instance, their sex, age, ticket class, fare, whether they had siblings or parents onboard and their cabin number. More details can be found on the data tab. Of course, we also know whether each passenger survived or not. The project description mentions a 38% survival rate.

Step2 - Prepare data Next, read in input data files into their respective pandas dataframes. Note that we have a training data set to train our model and a test dataset which will be used to validate our dataset. This can also be accomplished if one splits the training data up and reserves a portion of it for later.

Step 3 - At this point, we'll understand the position to perform the proper data and know which ML models are most likely to give the best results. Applying Pandas tools to have a first look at the data very easily.

trian\_df.info() & test\_df.info()

Step 4 - Perform basic statistics on numerical values only we observe. that the survival rate is the 38% mentioned earlier. Pclass has 3 integer values (also mentioned in description) we have some incomplete values for Age (714 non-null values of a total 891) and Cabin (only 204 non-null values)

**Step 5 -** Pivot features against Survival**.** Proceeded further into the Data Analysis. Correlation, Completing, Correcting, Creating, Classifying. Some patterns are starting to become evident from the pivot tables above - FirstClass passengers are more likely to have survived, as are women. SibSp and Parch need further investigation.

Step 6 - Data visualization Investigate a bit further using histograms. Plot the numerical features vs. survival We will use the Seaborn package (FacetGrid)

Observations 1

* Young Children had a high survival rate (>20%)
* Many 15-25 year olds did not survive.
* Majority of passengers are in the 15-35 age range.

## Observations 2

Pclass=3 had the most passengers but most of them did not survive. The youngest passengers (ie less than 5) in Pclass=2 and Pclass=3 mostly survived. Most passengers in Pclass=1 survived. Pclass varies in terms of Age distribution of passengers.

Based on these observations, **Pclass** is good to consider for model training

Observations 3

* PClass 1 and 2 generally had higher survival rates
* Females in 2 of the 3 embarkment points had significantly higher survival rates than males. A similar pattern is observed for the other embarkment point for males (C)
* Point of embarkment does have a varying effect on survival for males.

Based on this, Embarked and Sex will be added to model training. Embarked has 889 non null values so the remaining two will need to be completed.

Observations 4

* Embarkment point affects survival
* Again, in general females have higher survival rate
* Higher paying passengers had higher survival rate

Step 6 - Data Wragling - dropping Cabin and Ticket features. These should be done on both training and test data sets