### Kaggle - Titanic Machine Learning from Disaster

**Goal:** To predict whether a passenger survived or not

**Dataset provided**: train.csv, test.csv, gender\_submission.csv

**Target value**: Survived

Features: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked

Step 1]

* Import the necessary libraries & load the dataset into corresponding dataframes.
* Look upto 5 rows from train.csv file to find any relationship between features.

Step 2]

* It can be seen that train.csv consists of various numerical, categorical and ordinal features
* Datatypes in train.csv-> int, float & object datatype
* “Survived” feature tells us whether the passenger survived the fatal accident or not.
* There are many missing, misspelled values in train.csv and test,csv

Step 3]

* Visualize the columns using matplotlib & seaborn functions like pivot\_table, count\_plot, heatmaps, violinplots, histogram and bar charts
* Also, groupby the features with “Survived” feature to see their respective degrees of survival or accident.

Step 4]

* Detect & clean the dataset of any missing values. For eg, “Cabin” consisted of more than 75% missing values, so it would be wise to drop that column
* “Embarked” had a minimal missing instances, so we could impute the data with the most occurring instance.
* Similarly, “Age” has more missing data which can be meaned out

Step 5]

* Now, we will extract age ranges from “Age” column & convert these ranges to a numerical value
* Similar procedure was followed for “Fare”, so we have only int type values which are easier to process
* “Title” feature was extracted from “Name” feature
* “Family” & “IsAlone” feature was formed from “SibSp”(Siblings & Spouses) and “Parch”(Parents & children) feature
* Some additional feature: Age \* Pclass

Step 6]

* Keep only required features & drop the other features(do not drop PassengerId from test\_df) like SibSp, Parch, Survived(from train\_df), PassengerId

Step 7]

* Models:(Logistic Regression, Decision Tree, kNN, Random Forest)
* Before implementing model we assign X\_train, y\_train & X\_test
* A comparative dataframe was build to select the model with high accuracy

Step 8]

* K-fold cross validation was carried out to test the best model according to the folds
* Best feature importances were selected

Step 9]

* Prepare a submission.csv file to submit your predictions