**TITANIC PROBLEM STATEMENT**

Here, our problem statement is to predict whether the travelled passengers are likely to survive or not, based on the other feature columns in the data set. Below are the feature columns present in the dataset:

# survived: survival status of the person Survived - 1 Died - 0

# pclass : passenger class to which they belong

# sex: Male=1, Female=0

# age : age of the person

# sibsp: no of siblings & spouses

# parch: no of parents & children

# fare: ticket fare

# embarked: embark town

# class: passenger class

# who: sex of the passenger

# adult**\_**male: male passengers

# deck: deck name

# embark**\_**town: the town where the passenger got on

# alive **:** passenger alive or not

# alone: travelling alone or not

We have done exploratory analysis with the feature columns to understand their inter relationship and also their relationship with the target variable.

* Initially we studied the given columns & their interrelationship between them in the dataset.
* PARTITIONED THE DEPENDANT ( TARGET) COLUMN & THE INDEPENDENT (FEATURE) COLUMNS
* Dropped some of the unnecessary columns
* Checked the count of the different elements in “survived” column
* Checked for the presence of null values in the dataset.
* Used “parch” column to compute the missing values of “age” as “age & parch” columns are related to each other
* Plotted few plots including boxplot, heatmap, barplot to find some insights about the survival rates w.r.t sex, deck & age columns
* Found that female passengers survived more than male passengers
* FILLed THE MISSING VALUES OF DECK COLUMN WITH ITS MODE COLUMN
* Performed Label Encoding on “sex” column as it is nominal
* Performed One-Hot Encoding on “embark\_town” & “deck” columns as it had more than 2 values
* Now the titanic dataset values are transformed into 0’s & 1’s & replaced all the null values
* Dataset is cleaned now & it is ready for model building
* Then we divided the dataset into two variables, x\_data-having all the variables apart from the target variable, and y-having only the target variable using train\_test\_split
* Next step was to split the dataset in Train and Test and make predictions using the test data.
* Next, applied the models: Logistic Regression on trained data & noted down the accuracy, precision and recall score with respect to test and predicted values.
* **Conclusion**: The model had a high precision & low recall value which implies that it is likely to miss actual positive instances.