# EDA on Titanic Data Set

Hemalatha Subbiah

Bellevue University

DSC530-T303 Data Exploration and Analysis (2235-1)

Professor (or Dr.) Matthew Metzger

Date:06/03/2023

## Outcome of your EDA

I started with the data exploration where I got a feeling for the dataset, checked about missing data and learned which features are important. During this process I used seaborn and matplotlib to do the visualizations. During the data preprocessing part, computed missing values, converted features into numeric ones, grouped values into categories and created a few new features. Afterwards started training machine learning models, values. Lastly, looked at it’s confusion matrix and computed the model precision.

## What do you feel was missed during the analysis?

Not sure what methods need to be used for handling missing values in dataset. Out of this methods Deleting Rows with missing values, impute missing values for continuous variable, impute missing values for categorical variable, Other Imputation Methods, Using Algorithms that support missing values, Prediction of missing values I just used deleting Rows with missing values.I should have have used other smart ways to figure out the missing datas.May be tried replacing mean, median, or mode for Columns in the dataset which are having numeric continuous values.

**Were there any variables you felt could have helped in the analysis?**

The training set includes passengers’ survival status (also known as the ground truth from the titanic tragedy) which along with other features like gender, class, fare and pclass (passenger class) is used to create the machine learning model. The test set should be used to see how well the model performs on unseen data.

**Were there any assumptions you felt were incorrect?**

Completed age and sex feature as its correlated to Survival as multiple NA values were found. The ticket feature may be dropped from our analysis as it contains high ratio of duplicates (22%) and there may not be a correlation between Ticket and survival. PassengerId may be dropped from the training dataset as it does not contribute to survival. The upper-class passengers (Pclass=1) were more likely to have survived. SibSp feature has zero correlation for survival.

**What challenges did you face, what did you not fully understand?**

While building the machine learning models, the initial focus should be to select the feature set that is applicable for the target variable. In the case of the titanic dataset, the data pertains to routine passenger data collected which has nothing to do with accident or survivorship.

In the case of the titanic dataset, choosing some three variables viz., sex, parents, and children & siblings, we get an accuracy of 78.7%. Can we predict for any current luxury cruiser, the survivorship, based only on these three variables, with 78% accuracy? Can we get meaningful insights that are applicable to the entire population of luxury cruisers? The titanic dataset is simply not collected for this purpose.

## Conclusion

The conclusion I want to state is the titanic dataset I picked is to learn machine learning to beginners like me, promoted even by Kaggle. However, using the titanic dataset to teach and understand machine learning can lead to several misconceptions for beginners about the principles underlying the building of true machine learning models. Here are some important principles not emphasized using datasets like Titanic:1. Machine learning models are meant to solve business problems and help in decision-making. They are not meant to be used as a one-off exercise to gain insights. 2. Machine learning models are meant to be reproducible, replicable and applicable in similar contexts.