# Machine Learning Project Status Report

## Data Set Titanic Kaggle Data Set

### Dharmam Buch (DSB150030) and Darshan Parikh (DKP150130)

#### Introduction:

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. This sensational tragedy shocked the international community and led to better safety regulations for ships.

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 there 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.

#### Aim:

Complete the analysis of what sorts of people were likely to survive. In particular, apply the tools of machine learning to predict which passengers survived the tragedy.

#### Attributes:

1. Pclass – Passenger Class. (1 = 1st; 2 = 2nd; 3 = 3rd).
2. Name – Name.
3. Sex – Sex.
4. Age – Age.
5. Sibsp - Number of Siblings/Spouses Aboard.
6. Parch - Number of Parents/Children Aboard.
7. Ticket -Ticket Number.
8. Fare - Passenger Fare.
9. Cabin – Cabin.
10. Embarked - Port of Embarkation. (C = Cherbourg; Q = Queenstown; S = Southampton).

#### Class Attribute:

1. Survival – Survived or not ?(0=No, 1=Yes)

**Data Size (Number of Instances):**

1. Train Data – 891 Passengers.
2. Test Data – 418 Passengers.

Summary of data types after reading in R.

> str(d)

'data.frame': 891 obs. of 12 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Name : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...

$ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...

$ Embarked : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

**Training Methodology:**

* 1. **Initial Filtering**
  2. We will remove the Name from the training model, as the name is a continuous data and will never contribute to the survival of the person.
  3. Will also manually remove/exclude the Ticket data, as the Passenger\_class attribute is already there and ticket data in unreliable. For example a family of 3 will have one person with a high ticket and others with zero thus will remove this attribute as well.

1. **Deal With Nulls**
   1. Find columns with Null values, use the following R Command.

> colnames(d)[colSums(is.na(d)) > 0]

[1] "Age"

* 1. We will replace the empty age cells with the mean age of the data sets, hence, not destroying the mean of the data.

1. **Build model** 
   1. Using Random Forest and SVM.
   2. Use R studio with packages randomForest and e1701.
   3. If time permits use bootstrapping approach too.
2. **Validation**
   1. Use Model to predict the survived or not on the test data.
   2. Preliminary validation can be done by splitting the train data 80:20 and check the accuracy.
   3. Test Accuracy will be provided by kaggle website as the challenge is still on.
   4. Once the test results are provided by kaggle website and time permits, I would like to do empirical studies on accuracy of different models on these data sets.