# Introduction

In this paper, we will use the famous dataset “Titanic: Machine Learning from Disaster” from Kaggle.com to explore the feature importance to analyze survivability based on passenger data. By the end of this project, we will build a machine learning model to predict which passenger survived the Titanic Shipwreck.

For this project, we will use ensemble machine learning techniques like random forest and Gradient boosting to understand the working of the ensemble learning method and we will finally compare the result of the ensemble model with conventional algorithms like logistics regression and Decision tree.

The goal of this project is to build a model to predict if the passenger survives the disaster or not, with the highest possible accuracy score. To achieve high accuracy, we will try to extract as much data and information we can, from the given dataset using feature engineering. Other challenges in this dataset include cleaning and missing values. There 866 missing values in this dataset that needs to be adjusted.

## Dataset and Data Dictionary

The given dataset contains records of 891 passengers onboard on Titanic, the dataset includes 12 features including passenger ID. The given dataset is only the training part from the Kaggle challenge, however, for this project, we will use the given set as the complete set and later split into training and testing set. The dataset includes passenger information like name, age, gender, soci0-economic class, ticket id, number of family members, ticket fare, etc. And most importantly, the dataset reveals if each passenger survived or not.

Data Dictionary:

Survival : 1 if passenger survived, 0 if not (binary)

Pclass : 1st, 2nd, and 3rd class of Ticket (int)

Sex : Gender of passenger (binomial string)

Age : Age in years (int)

Sibsp : Number of siblings and/or spouses on board during the disaster(int)

Parch : Number of parents and/or children on board during the disaster(int)

Ticket : Ticket number (unformatted string)

Fare : Passenger fare (float)

Cabin : Cabin number (an alphanumeric string)

Embarked : Port of Embarkation, C for Cherbourg, Q for Queenstown, and S for Southampton

# Analysis and Interpretation

We will use pandas and NumPy for data manipulation and data cleaning. To train the model and test its accuracy, we will be using different libraries from Sci-kit learn (sklearn) package. Along with that, we will also use matplotlib and seaborn for data visualization for the ease of exploratory data analysis.

We will start by loading all the libraries that will be used in this project.



We will load the given dataset of dimension 891 \* 12 into a data frame object called ‘data’. There are total of 866 missing value in the data frame, we will check number of missing values in each column to decide treatment on missing value.

Survived 0

Pclass 0

Name 0

Sex 0

Age 177

SibSp 0

Parch 0

Ticket 0

Fare 0

Cabin 687

Embarked 2

From above output, we can see that huge pool of missing values are from “cabin” column followed by “Age” and “Embarked”. Since, the number of missing values is huge, we will have to find a way to fill the empty cell. Also, the initial hypothesis based on the overlook of this dataset is that the survival of passenger is hugely influenced by their gender, age and class. We will find way to fill missing value on each column.

## Data cleaning and Feature extraction

Filling missing values in each column by:

1. **Embarked**: Only two value are missing in Embarked column, we will fill the missing value with Embarked class with highest mode. ‘S’ has highest frequency of 644 in 889 records, hence **we are going to fill null in Embarked with ‘S’**.
2. **Cabin**: Cabin is alphanumeric string with alphabet, assuming its initial alphabet as identifier of which cabin passenger was assigned, we will only extract cabin alphabet as from cabin number.

Pivoting Class against Cabin alphabet, we can see that only 9 passengers with 3rd class ticket are assigned cabin, this indicates that majority of cabin number missing values are passenger without cabin, who were allotted general space in ship. We can use this feature to later analyze that how likely were passengers without cabin would survive compared to passenger with cabin.

**We will replace null values in this column by string “No cabin”.**

1. **Age:** There are 177 missing values in “Age” column, we can directly fill the null value with average age of available data. However, based on our hypothesis, age is going to be important deciding factor to predict survivability of passenger, as it is expected to prioritize women and children.

We will take few steps of feature engineering which will later help us fill the missing value of age. We will extract the title of each passenger from “Name” column. Multiple different titles were found, so we will limit the title categories to the one with more than 10 frequency, title with less than 10 counts will be labeled “Others”.

Now that we have 5 major titles Mr., Miss, Mrs., Master and Others, **we will fill missing age data with average age of respective title categories.**

From above steps we have cleaned data and extracted two important feature, Cabin allocation(Yes or no), if yes, which cabin as an alphabet indicator and second and important feature, Title of Passenger.

## Modeling

As part of modelling, we will first convert each of the categorical variable to numerical using one hot encoding, we are doing this transformation to adjust with compatibility of sklearn modeling functions. To avoid multicollinearity, one of each dummy variable have been removed. Also, few columns from original dataset have been removed as we have already extracted important feature form those columns.

Next step is to split the given dataset into training and testing set, we will split in the ratio of 7:3. Also, we will scale all the feature variable because few of the variable such as age and fare have huge range compare to others.

As discussed above, to compare the prediction accuracy of different modeling technique, we will be fitting multiple models. We will also model using ensemble learning which is combination of multiple base models.

Below is the accuracy of 4 models when tested with testing set which was not used to train the model:

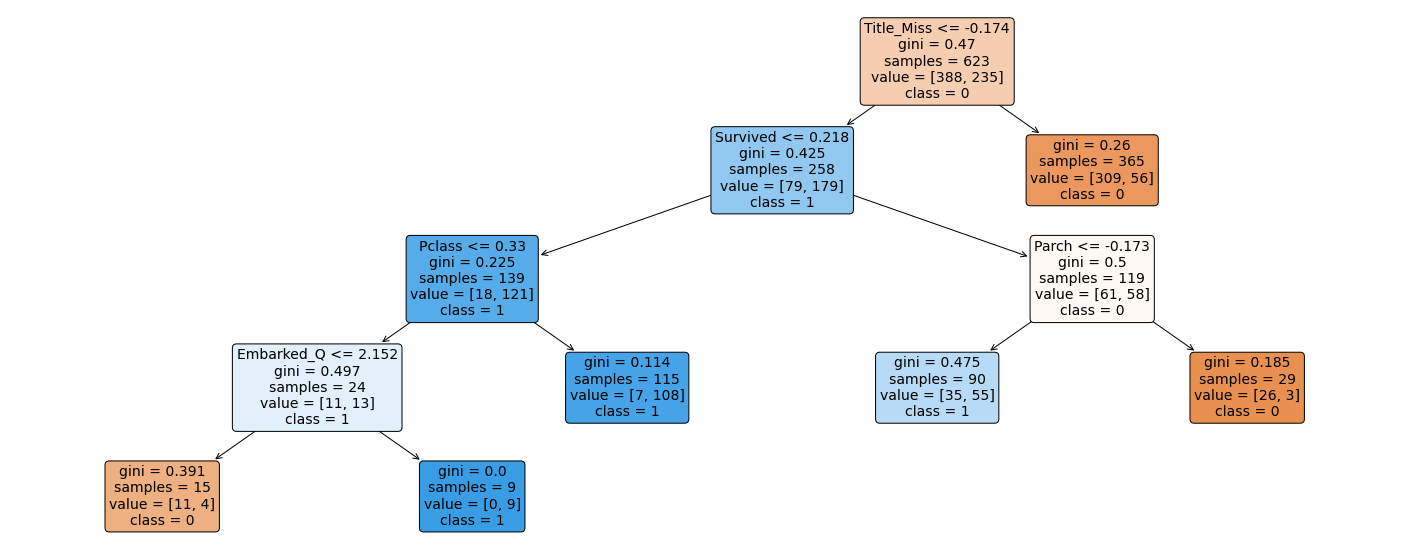
**Model-1 (Decision Tree): 83.58%**

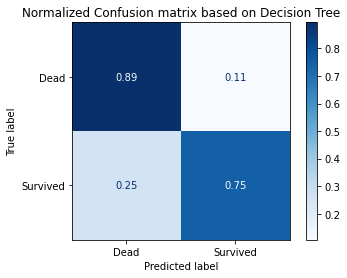
**Model-2 (Random Forrest): 79.47% (Ensemble)**

**Model-3 (Logistic regression) : 83.20%**

**Model-4 (Gradient Boosting): 82.83% (Ensemble)**

Based on the result above, we can see that Decision tree method delivered better result followed by Logistics regression. Ensemble learning is expected to do a better job as it is combination of multiple base models but is not correct in this case because our target variable is categorical data. In accuracy in Classification problem is a huge disadvantage of ensemble learning technique, however they are considered good for regression modeling.





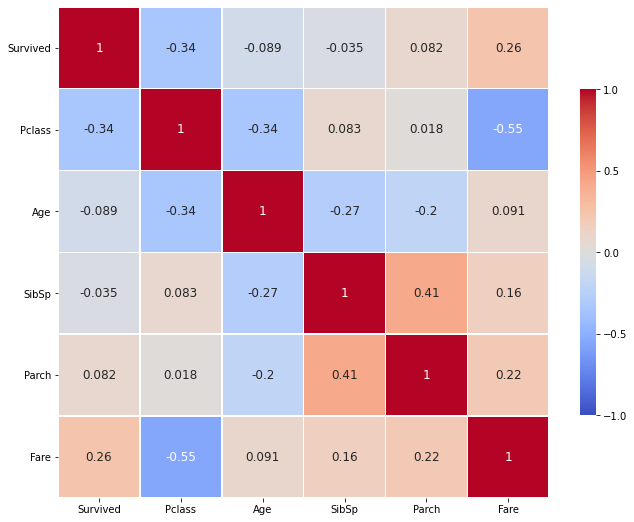
The above Decision Tree uses Gini index as a classifier to branch out decision. Gini index is impurity measure which when 0.5 means classes are distributed equally. Decision tree above is visualization of fitted model based on this class is decided for each record.

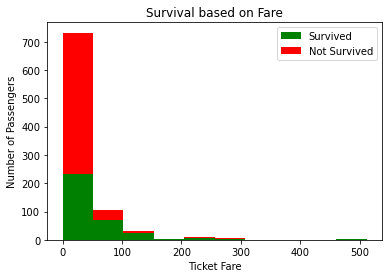
Based on Normalized confusion matrix, decision tree was able to predict 89% of True Negative correctly and 75% of True Positive correctly. The overall accuracy of the above decision tree model is 83.58%.

# Conclusion

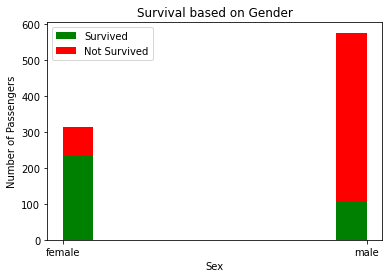
In conclusion section, we will discuss what feature were high influence on the survivability of passenger in shipwreck of Titanic.

We will start by looking correlation between numerical variables.

From this correlation coefficient visualization between variable, we can see that Ticket class has strong negative correlation with survival. “Pclass” in this class is original where 1 indicate first class, this means the passenger with higher class ticket were more likely to make it out alive than lower class. Similar indication can be observed in fare, fare has strong positive correlation with survival, that mean higher the fare, more than chance of survival. Similarly, higher number of sibling or spouse mean low chance of survival.

We will explore these finding from correlation coefficient using different visuals.

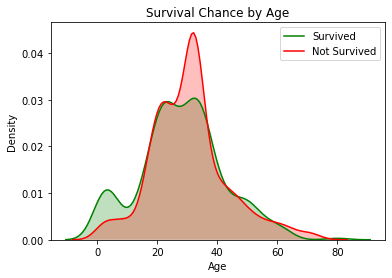
From this graph, we find that most of the ticket are less priced and as suggested by correlation coefficient higher proportion of cheap fare passenger could not survive. In diagram, we can notice very small proportion of passenger with fare between 50 to 100 and 100 to 200 were not rescued, this suggest rescue operation was also based on class.



Based on visuals below, we can notice that there were twice male passengers compared to females, however only less than 20% male survived, but almost 80% females survived. This suggests that females were prioritize for rescue ships.

As we initially hypothesized that, women and children would have been the priority for rescuers, it is found to be true for women.

Now we will visualize the age distribution along with the chance of survival to understand if there was any rescue priority based on age of passenger.

The density plot of passenger record, passenger who survived and passenger who did not alongside can helps us understand if there was priority based on age.

Based on this diagram, there was clearly priority for kids, skewness in green curve indicated, survival of passenger below 20 is more. Similarly, skewness in red curve can be seen between age 25 to 40, this mean most of the passenger who did not make it are from this age group. However, for citizen for than 40 years old, there seems to be almost equal possibility of survivability.

|  |  |
| --- | --- |
|  |  |
|  |  |

Above, probability charts visualize chances of survival for male and female in different categories. As expected, survival probability decreases with the increase in number of siblings, spouse or parents onboards. We can also see that, the probability of survival for female in cabin A, B, D and F is 100%, which means these cabins were rescued first, although female have 100 survivability in these cabin, males have less compared to cabin E. Most significantly, we can observe that probability of survival was quite low for passenger with No cabin, both male and female. We can also notice that something unusual happen on cabin G, either it was in difficult part of ship which are directly impacted by disaster or it was last served in rescue operation. No men from cabin G survived and only 50% females were rescued, which is quite less compared to other cabin, even lesser than No cabin passenger.

# Reference

Kaggle(2012), Titanic: Machine Learning from Disaster, Retrieved from <https://www.kaggle.com/c/titanic/overview>

Scikit-Learn (n.d), Documentation (multiple), Retrieved from <https://scikit-learn.org/stable/index.html>