# **ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to my Mini Project Supervisor **Dr. Arun Kumar Singh** as well as our HoD **Prof P.C Vashist** who gave me the golden opportunity to do this wonderful project on “**EXPLORATORY DATA ANALYSIS ON TITANIC DATASET”** and for their intellectual, valuable guidance which helped me a lot in doing my work. This project would have been their enormous help and worthy experience. Whenever I was in need, they were there behind me. It makes me to do a lot of research work by which I came to learn about so many new technologies.

Although, this project has been prepared with utmost care and deep routed interest. Even then it was not possible to complete it without any support.

During this period I’ve come a long way in using statistics and data visualization to understand relationships among variables in the data. I’m glad to admit that I feel comfortable using Statistics to understand patterns in data. I’m looking forward to using more advanced statistical methods to derive insights from more complex data with entangled variables.

Last but not the least, I would also extend my gratitude to my parents and friends who helped me a lot in finalizing this project within the limited time frame.

# **SYNOPSIS**

In this project, We'll be trying to do data analysis on titanic dataset and predict a classification - survival or deceased using Logistic Regression. We'll use a "semi-cleaned" version of the titanic data set present on largest machine learning learning platform named kaggle.

The most infamous disaster which occurred over a century ago on April 15, 1912, that is well known as sinking of “The Titanic”. The collision with the iceberg ripped off many parts of the Titanic. Many classes of people of all ages and gender where present on that fateful night, but the bad luck was that there were only few life boats to rescue. The dead included a large number of men whose place was given to the many women and children on board. The men travelling in second class were dead on the vine.

Machine learning algorithms are applied to make a prediction which passengers survived at the time of sinking of the Titanic. Features like ticket fare, age, sex, class will be used to make the predictions. Predictive analysis is a procedure that incorporates the use of computational methods to determine important and useful patterns in large data. Using the machine learning algorithms, survival is predicted on different combinations of features.

The objective is to perform exploratory data analytics to mine various information in the dataset available and to know effect of each field on survival of passengers by applying analytics between every field of dataset with “Survival” field. The predictions are done for newer data sets by applying machine learning algorithm. The data analysis will be done on applied algorithms and accuracy will be checked. Different algorithms are compared on the basis of accuracy and the best performing model is suggested for predictions .

Steps involved in Exploratory data analysis (EDA) :

1. Import Dataset & Headers
2. Identify Missing Data
3. Replace Missing Data
4. Evaluate Missing Data
5. Dealing with Missing Data
6. Correct Data Formats

vii. Data standardization

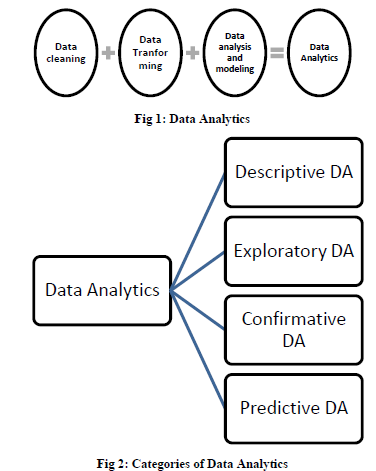
viii. Data Normalization

ix. Binning.

x. Indicator variable.

**Key Words** **Used**: Logistic Regression, Data Analysis , Kaggle Titanic Dataset, Data pre-processing . Cross validation, Confusion Matrix

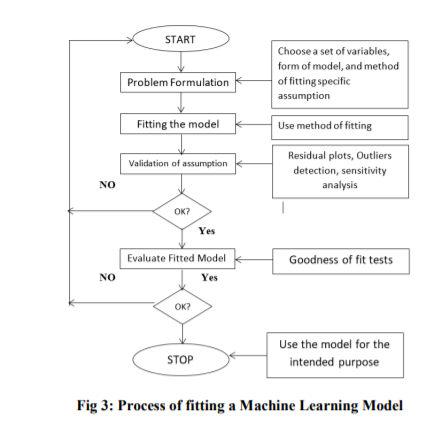
**1) DATA ANALYTICS AND ITS CATEGORIES** :



**Fig 2 – Categories of Data analysis**

**2) PROCESS FLOW :**

There is a step by step approach to choose a particular model for the current problem. We need to decide whether a particular machine learning model is suitable for our problem or not. Here we can see process flow being followed



**Fig 3: Process fitting**

**3. ALGORITHM :**

**3 - (a) - DATA PREPROCESSING:**

In the dataset available for the prediction some of the data values are missing or unknown. This missing data was resulting in reducing the accuracy of the overall prediction model and also reduces the size of pure training data which in turn reduces accuracy. Data preprocessing is a technique that involves transforming raw data into an understandable format.

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Missing values are replaced by average of that column. So, the missing and unknown data of the passengers which is easily predictable is filled up by this step.

**3) - (b) – CLASSIFICATION ALGORITHM I.e**

**LINEAR REGRESSION :**

Second step of the algorithm is using a classifier to classify the available information. Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.

Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It uses a method of using he regression line between dependent and independent variable to predict the value of the dependent variable

**3) - (c) - CROSS VALIDATION :**

Dataset is divided into two main parts namely Train and Test data. Training data will be considered for the training of the machine. Test data will be used for validating the machine. Cross validation technique used here is K-Fold.

The method has only one parameter called k that refers to the number of groups into which a given data sample is to be split. As such, the method is also called k-fold cross validation. When a particular value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Analysis of confusion matrix Confusion matrix is used to show the performance of the algorithm. Accuracy of the model can be predicted using the confusion matrix. It is a plotting of relation between real and predicted outputs. It allows us to check the accuracy and performance of the algorithm. In this case we are using two attributes at a time for the confusion matrix plotting. Test case data is used to build the confusion matrix. The values shown in the confusion matrix are the probability of survival of the individual considering only those parameters.

**3) – (d) -.RESULTS :**

The logistic regression gives the accuracy of 95% which is based on the confusion matrix. The parameters used here are accuracy and false discovery rate. Accuracy is a measure of the correctness of the prediction of the model. Higher accuracy is always better and is calculated by :

**(TN + TP)/Total number of rows \*100**

False discovery rate are the false positive measures of confusion matrix where the model predicts that the passenger would survive but in reality, it doesn’t. This would prove dangerous as the prediction may go wrong and hampers the accuracy of the results.

The attempts are being made to increase the accuracy rate and reduce the false discovery rates.

## **4) DATASET INFORMATION:**

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.

Titanic data  - Contains demographics and passenger information from 891 of the 2224 passengers and crew on board the Titanic.

| **Variable** | **Definition** | **Key** |
| --- | --- | --- |
| Survived | Survival | 0 = No, 1 = Yes |
| Pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| Sex | Sex |  |
| Age | Age in years |  |
| Sibsp | # of siblings / spouses aboard the Titanic |  |
| Parch | # of parents / children aboard the Titanic |  |
| Ticket | Ticket number |  |
| Fare | Passenger fare |  |
| Cabin | Cabin number |  |

**Pclass**: A proxy for socio-economic status (SES)  
1st = Upper  
2nd = Middle  
3rd = Lower

**Age**: Age is fractional if less than 1.

**Sibsp**: The dataset defines family relations in this way...  
Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored)

**Parch**: The dataset defines family relations in this way...  
Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny, therefore parch=0 for them.

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| SR. NO | CONTENT | Page No |
| 1 | ACKNOWLEDGEMENT |  |
| 2 | SYNOPSIS |  |
| 3 | STORYTELLING |  |
| 4 | QUICK GLANCE ON DATA |  |
| 5 | NUMERICAL VARIABLES |  |
| 6 | DATA DISTRIBUTION |  |
| 7 | IMPLEMENTATION FOR PREDICTING ACCURACY |  |
| 8 | CONCLUSIONS |  |
| 9 | REFRENCES |  |

**INTRODUCTION**

The sinking of the Titanic ship caused the death of about thousands of passengers and crew is one of the fatal accidents in history. The loss of lives was mostly caused due to the shortage of the life boats. The mind shaking observation came out from the incident is that some people were more sustainable to endure than many others, like children, women were the one who got the more priority to be rescued. The main objective of the algorithm is to firstly find predictable or previously unknown data by implementing exploratory data analytics on the available training data and then apply different machine learning models and classifiers to complete the analysis.

This will predict which people are more likely to survive. After this the result of applying machine learning algorithm is analyzed on the basis of performance and accuracy

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modeling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data. Through the process of EDA, we can ask to define the problem statement or definition on our data set which is very important.

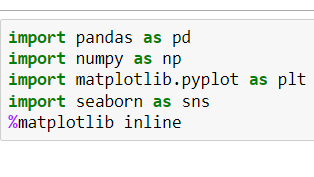
EDA is used :

* To give insight into a data set.
* Understand the underlying structure.
* Extract important parameters and relationships that hold between them.
* Test underlying assumptions

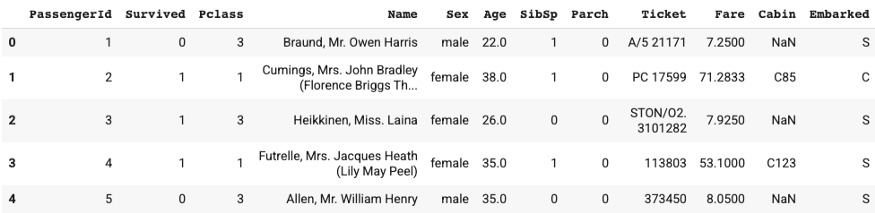
It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

**1)** **A quick glance on data :**

First, we will import the necessary packages and load the data set.

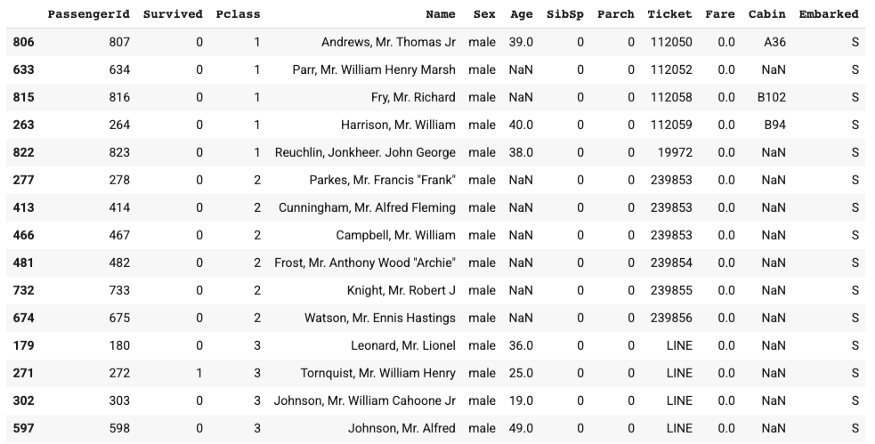


--------------------------------------------------------------



**Fig1 – Glance on data**

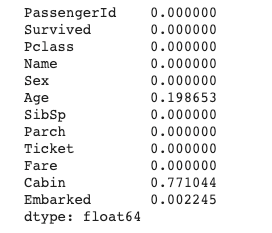
In the train data, there’re 891 passengers, and the average survival rate is 38%. Age ranges from 0.42 to 80 and the average is ~30 year old. At least 50% of passengers don’t have siblings / spouses aboard the Titanic, and at least 75% of passengers don’t have parents / children aboard the Titanic. The fare varies a lot.



**Fig2 – Train data**

Above is a list of passengers with $0 fare. We spot checked a few passengers to see if the $0 fare is intended.

Passengers that share the same ticket number seem to be in the same traveling group. We can create a boolean variable for traveling group to see if people travelled in groups would be more likely to survive.

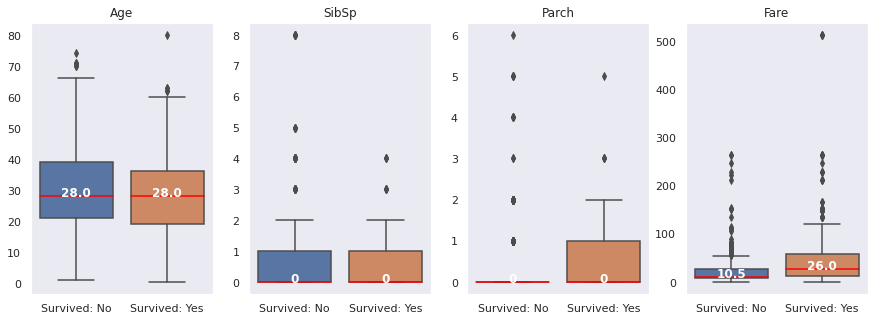


**Fig3 – Missing data**

20% of Age data is missing, 77% of Cabin data is missing, and 0.2% of Embarked data is missing. We’ll need to handle the missing data before modeling. This will be covered in Feature Engineering article as well.

**2) Numerical Variables:**

As to the box plots, survivors and victims have similar quartiles in Age and SibSp. Compared to victims, survivors were more likely to have parents / children aboard the Titanic and have relatively more expensive tickets.

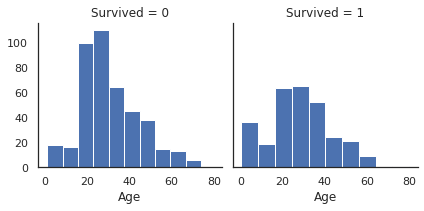
Box plot provides a quick view of numerical data through quartiles. Let’s also check the data distribution using histograms to uncoveadditionalpatterns.

**Fig**

As to the box plots, survivors and victims have similar quartiles in Age and SibSp. Compared to victims, survivors were more likely to have parents / children aboard the Titanic and have relatively more expensive tickets.

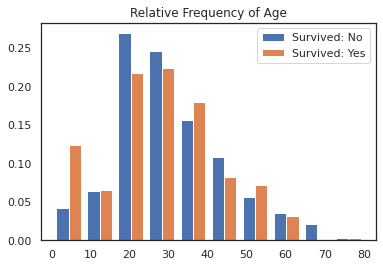
Box plot provides a quick view of numerical data through quartiles.

3) **Data disrubution :**



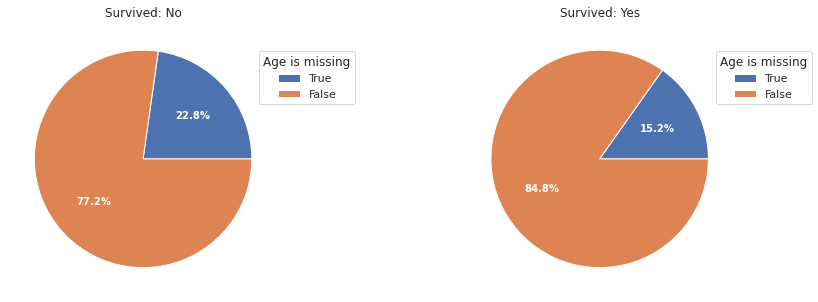
**Fig 5 – Distribution plot**

When comparing the distribution of two sets of data, it’s preferred to use the relative frequency instead of the absolute frequency. Using Age as an example, the histogram with absolute frequency suggests that there were a lot more victims than survivors in the age group of 20–30 .



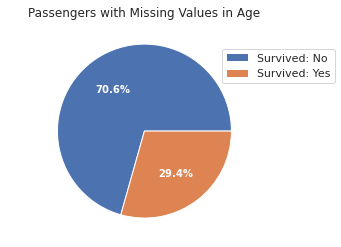
**Fig6 – Relative Frequency of age**

In the histogram of relative frequency for age, what really stands out is the age group < 10. Children were more likely to survive compared to victims among all age groups.



**Fig 7 – Pie Plot for Survived data**

From the pie plots, we can tell that passengers with missing age were more likely to be victims.



**Fig8 – Pie plot for missing age**

Regarding feature engineering for Age, I’ll probably create a categorical variable including categories for Children, Adult, Senior and Missing Values respectively.

**STORYTELLING**

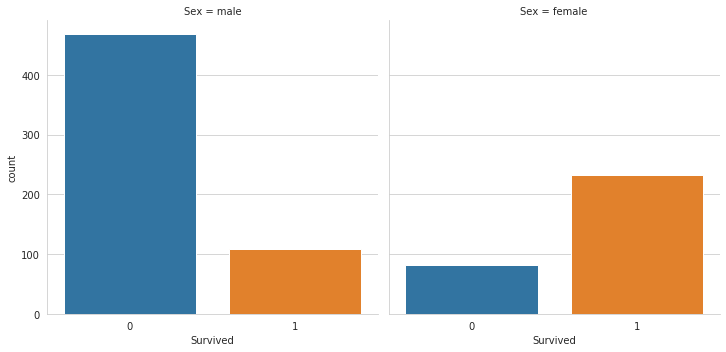


**Fig 9 – Null values**

The column ‘Age’ and ‘Cabin’ have got null values. While ‘Cabin’ has huge amount null values, ‘Age’ has moderate amount of null values.

We need to form a logic to impute the missing values of the ‘Age’ column. We shall come back to it later after understanding the relation between ‘Age’ and various other variables.

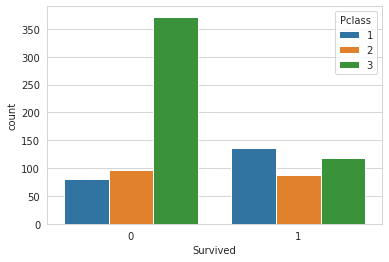
Let us try to know if the dependent variable ‘Survived’ has any relation with the variable ‘Sex’. To do so we would use factor plot.



**Fig 10 – Factorplo****t**

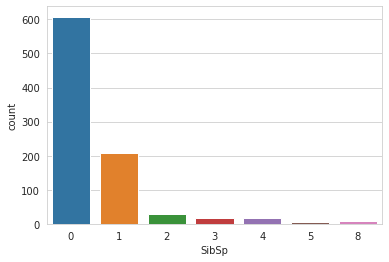
**Inference**: As we all know from the movie as well as the story of titanic females were given priority while saving passengers. The above graph also tells us the same story. More number of male passengers have died than female ones.

Similarly let us try to see how the variable ‘***Pclass***’ is related to the variable ‘Survived’.



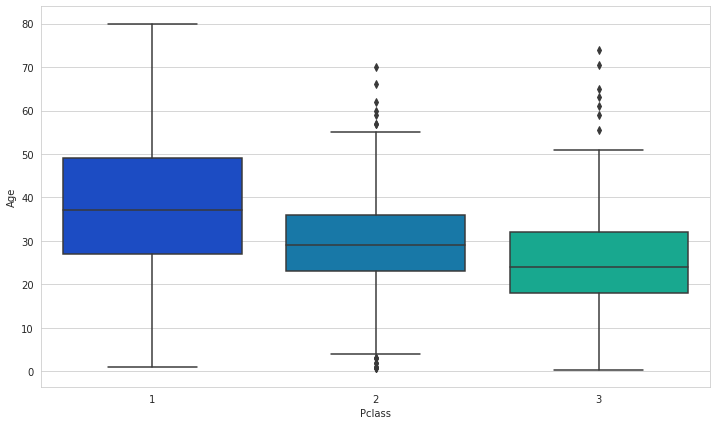
**Fig11 – Plot to find victim according to class**

The graph tells us that Pclass 3 were more likely to be survived. It was meant for the richer people while Pclass 1 were the most likely victims which was relatively cheaper than class 3.



**Fig12 - number of Sibling or spouse**

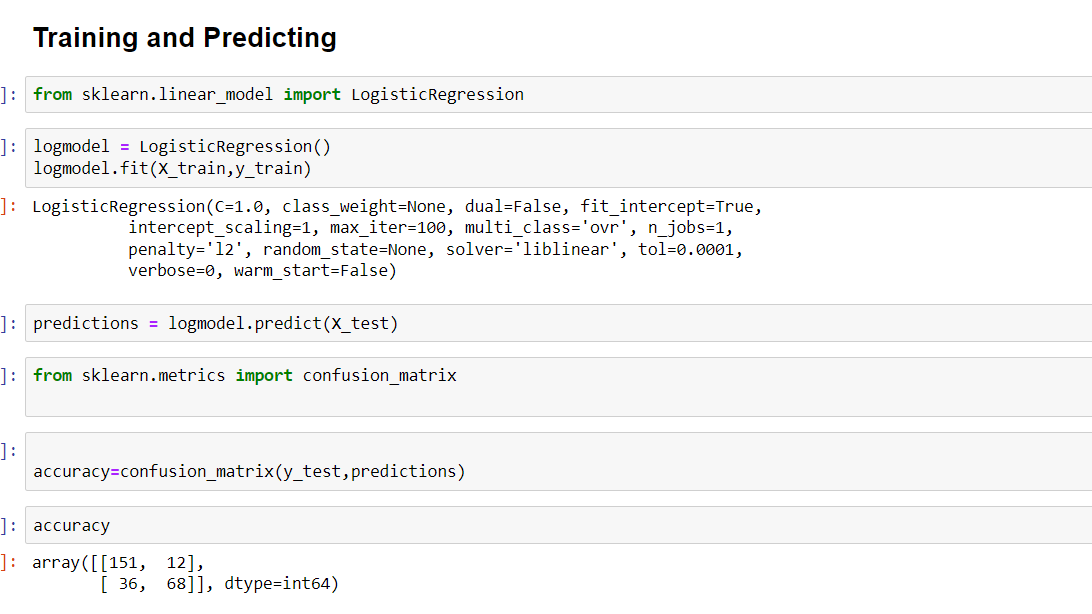
Here ‘SibSp’ variable refers to the number of sibling or spouse the person was accompanied with. We can see most of the people came alone.

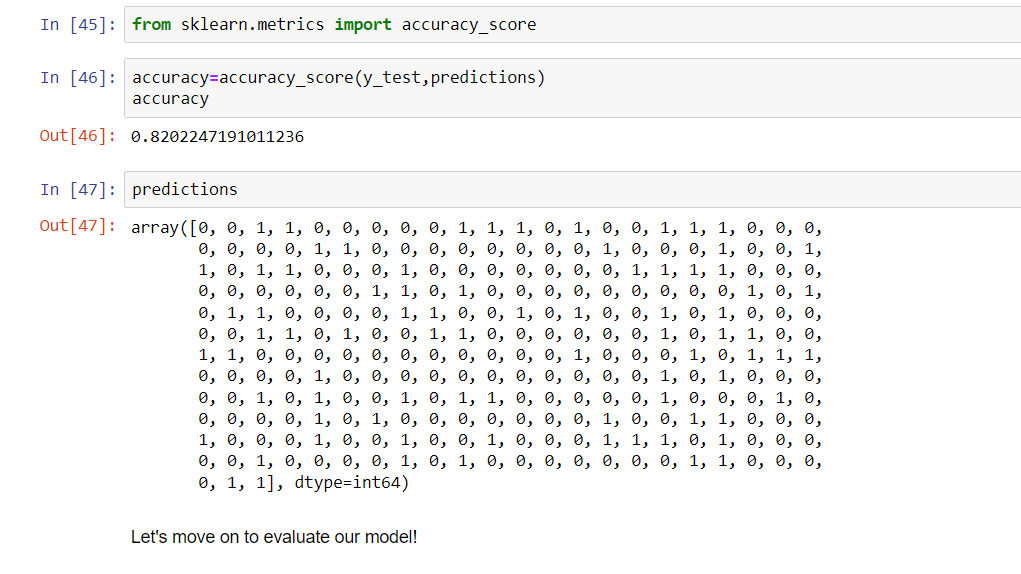


**Fig13 - Boxplot**

Now, figure out a way to fill the missing value of the variable ‘Age’. Here we segregated the ‘Age’ variable according to the Pclass variable as it was found out that ‘Age’ and ‘Pclass’ column were related. We would draw a boxplot that would tell us the mean value each of the Pclass.

**IMPLEMENTATION FOR PREDICTING ACCURACY**

****

******

***HENCE , ACCURACY OF THE PREDICTION = 0.82 i.e 82%***

# **CONCLUSION**

The logistic regression provides a better accuracy i.e. almost of about 82%. It works better with binary dependent variable which means the variable has a binary value as its output like yes or no, true or false.

In conclusion, we can say that this data gives us the information of the travellers and whether they survived or not.

The confusion matrix gives the accuracy of all the models, the

logistic regression is proves to be best among all with an

accuracy of 0.827261504. This means the predictive power of

logistic regression in this dataset with the chosen features is

very high.

It is clearly stated that the accuracy of the models may vary

when the choice of feature modelling is different. Ideally

logistic regression and support vector machine are the models

which give a good level of accuracy when it comes to

classification problem.

I really hope this has been a great read and a source of inspiration to develop and innovate.

# **REFRENCES**

[1]Analyzing Titanic disaster using machine learning algorithms-Computing, Communication and Automation (ICCCA), 2017 International Conference on 21 December 2017, IEEE.

[2]Prediction of Survivors in Titanic Dataset: A Comparative Study using Machine Learning Algorithms, Tryambak Chatterlee, IJERMT-2017.

[3]MICHAEL AARON WHITLEY, using statistical learning to predict survival of passengers on the RMS Titanic by Michael Aaron Whitley, 2015

[4]Atakurt, Y., 1999, Logistic Regression Analysis and an Implementation in Its Use in Medicine, Ankara University Faculty of Medicine Journal, C.52, Issue 4, P.195, Ankara

[5]MICHAEL AARON WHITLEY, using statistical learning to predict survival of passengers on the RMS Titanic by Michael Aaron Whitley, 2015.

[6]Bircan H., Logistic Regression Analysis: Practice in Medical Data, Kocaeli University Social Sciences Institute Journal, 2004 / 2: 185- 208

[7]Atakurt, Y., 1999, Logistic Regression Analysis and an Implementation in Its Use in Medicine, Ankara University Faculty of Medicine Journal, C.52, Issue 4, P.195, Ankara