A MINI PROJECT REPORT

EXPLORATORY DATA ANALYSIS ON TITANIC DATASET

Submitted by

NISHIT CHAUDHARY - 1901920130119
PANKAJ SHARMA - 1901920130120
PIYUSH SHARMA - 1901920130121

Submitted to

MR. ANAND BHUSHAN PANDEY
(Assistant Professor GLBITM Greater Noida)

Department of Information Technology

TABLE OF CONTENTS

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.

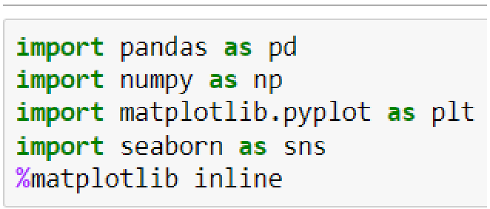
EDA is used :

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

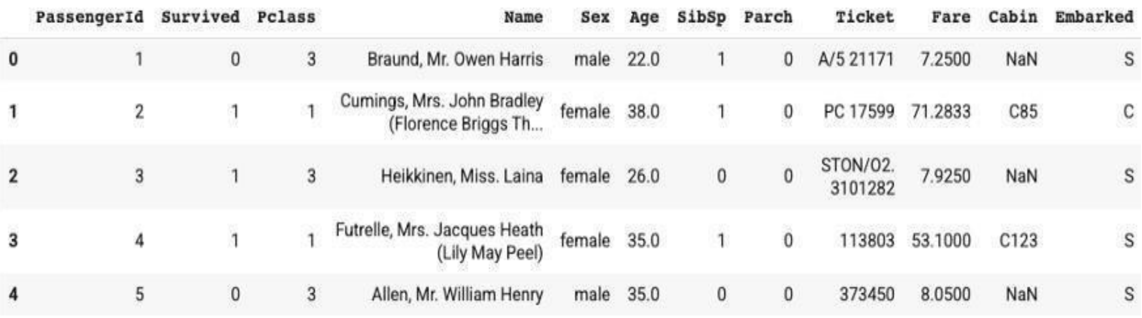
It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

1) A quick glance on data :

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



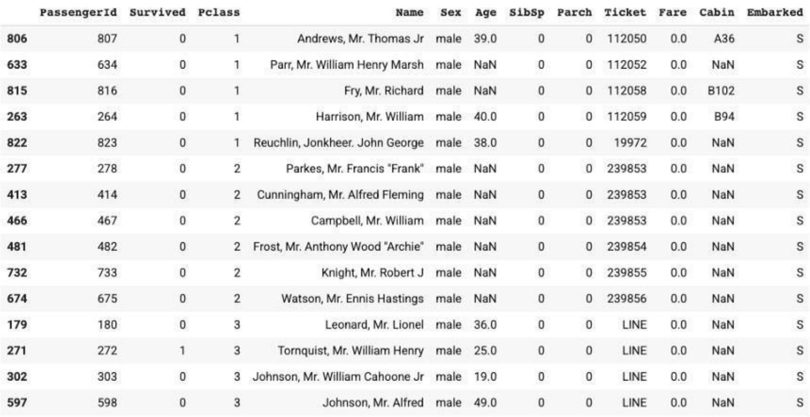
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Fig 1 : 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.

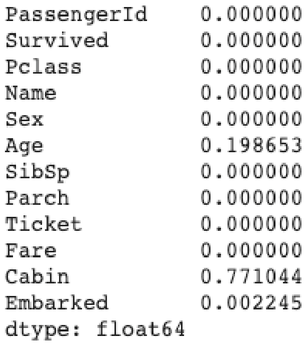


	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
806	807	0	1	Andrews, Mr. Thomas Jr	male	39.0	0	0	112050	0.0	A36	S
633	634	0	1	Parr, Mr. William Henry Marsh	male	NaN	0	0	112052	0.0	NaN	S
815	816	0	1	Fry, Mr. Richard	male	NaN	0	0	112058	0.0	B102	S
263	264	0	1	Harrison, Mr. William	male	40.0	0	0	112059	0.0	B94	S
822	823	0	1	Reuchlin, Jonkheer. John George	male	38.0	0	0	19972	0.0	NaN	S
277	278	0	2	Parkes, Mr. Francis "Frank"	male	NaN	0	0	239853	0.0	NaN	S
413	414	0	2	Cunningham, Mr. Alfred Fleming	male	NaN	0	0	239853	0.0	NaN	S
466	467	0	2	Campbell, Mr. William	male	NaN	0	0	239853	0.0	NaN	S
481	482	0	2	Frost, Mr. Anthony Wood "Archie"	male	NaN	0	0	239854	0.0	NaN	S
732	733	0	2	Knight, Mr. Robert J	male	NaN	0	0	239855	0.0	NaN	S
674	675	0	2	Watson, Mr. Ennis Hastings	male	NaN	0	0	239856	0.0	NaN	S
179	180	0	3	Leonard, Mr. Lionel	male	36.0	0	0	LINE	0.0	NaN	S
271	272	1	3	Tornquist, Mr. William Henry	male	25.0	0	0	LINE	0.0	NaN	S
302	303	0	3	Johnson, Mr. William Cahoone Jr	male	19.0	0	0	LINE	0.0	NaN	S
597	598	0	3	Johnson, Mr. Alfred	male	49.0	0	0	LINE	0.0	NaN	S

Fig 2 : 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.



```
PassengerId    0.000000
Survived       0.000000
Pclass         0.000000
Name           0.000000
Sex            0.000000
Age            0.198653
SibSp          0.000000
Parch          0.000000
Ticket         0.000000
Fare           0.000000
Cabin          0.771044
Embarked       0.002245
dtype: float64
```

Fig 3 : 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 uncover additional patterns.

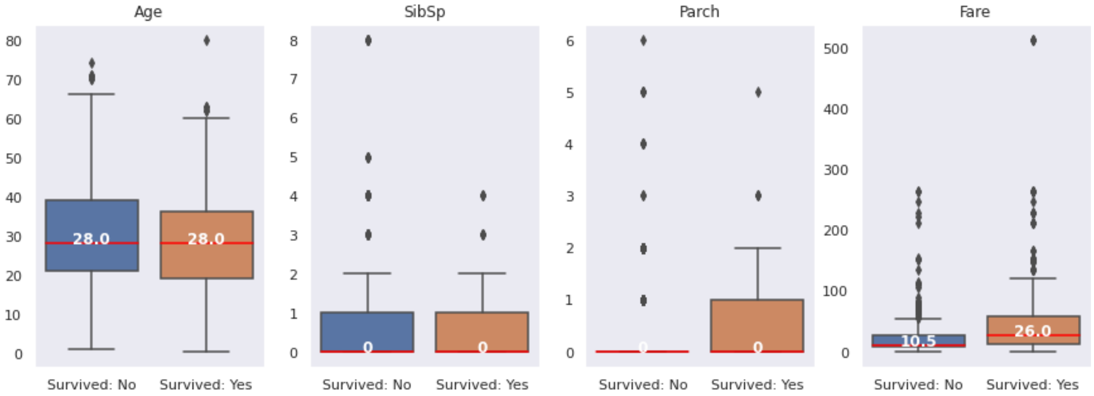


Fig 4 :Box Plot

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 :

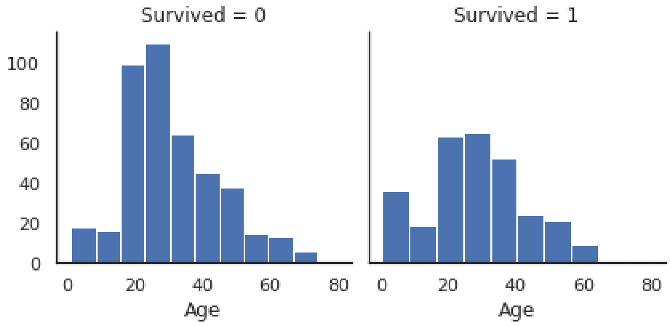


Fig 5 : Distribution plot

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 .

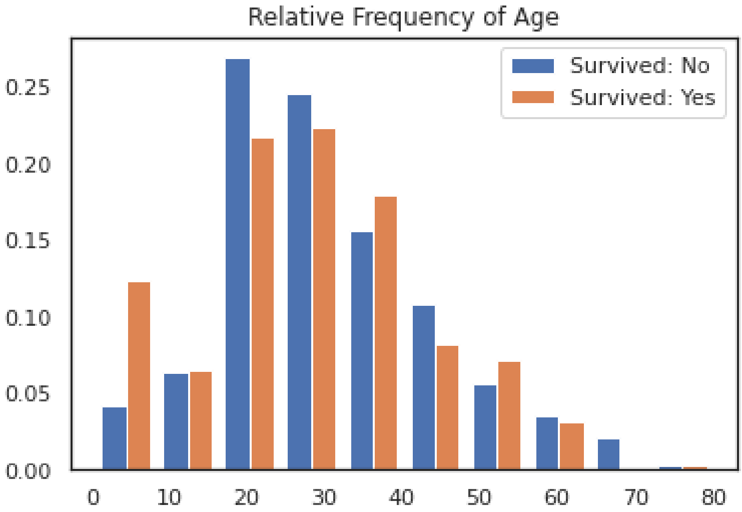


Fig 6 : Relative Frequency of age

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.

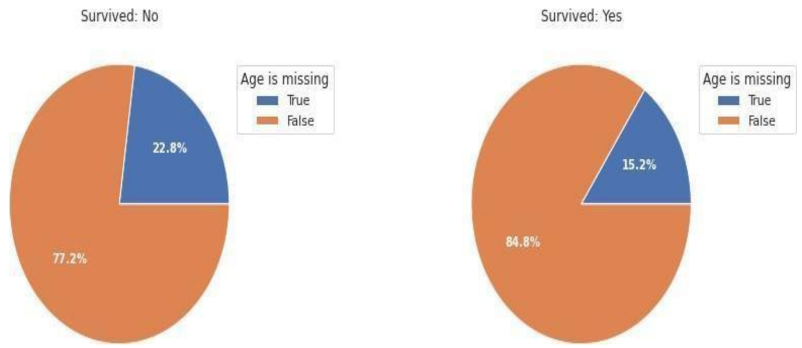


Fig 7 : Pie Plot for Survived data

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

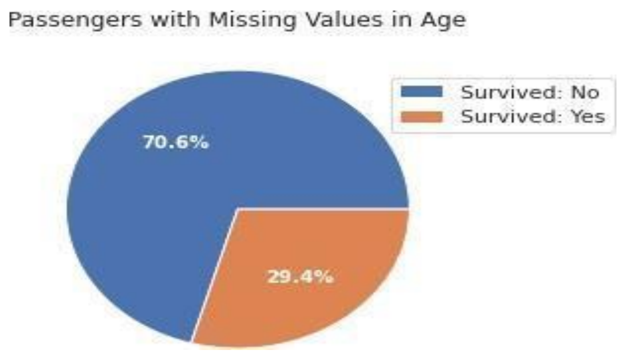


Fig 8 : Pie plot for missing age

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

STORYTELLING

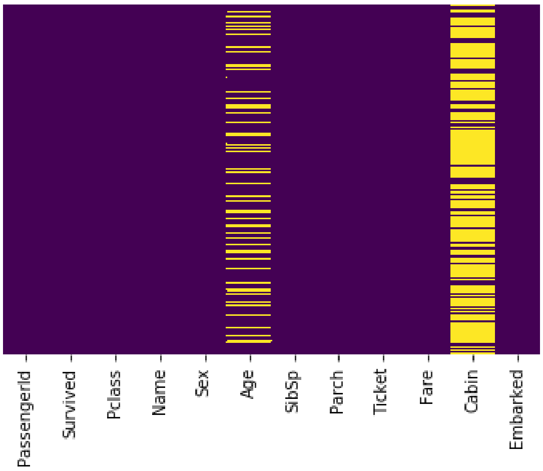


Fig 9 : Null values

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.

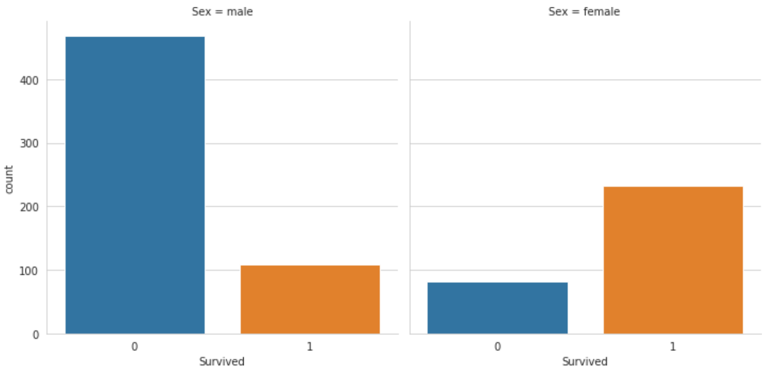


Fig 10 : Factorplot

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".

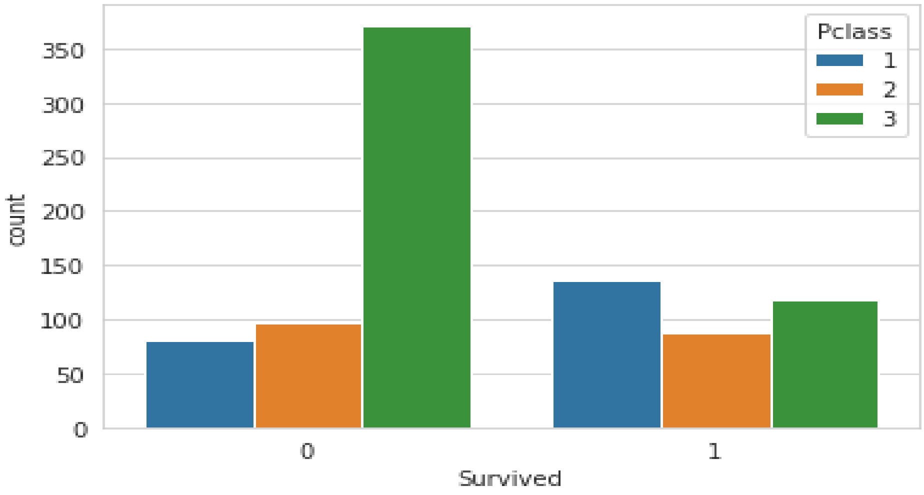


Fig 11 : Plot to find victim according to class

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.

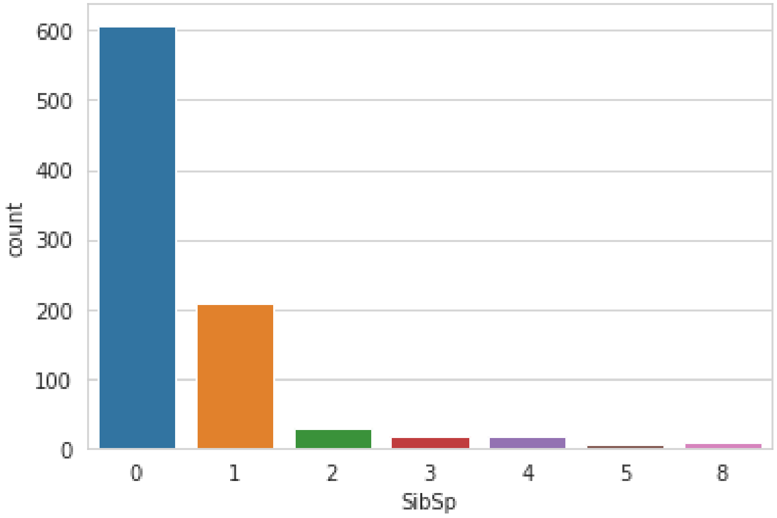


Fig 12 : Number of Sibling or spouse

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.

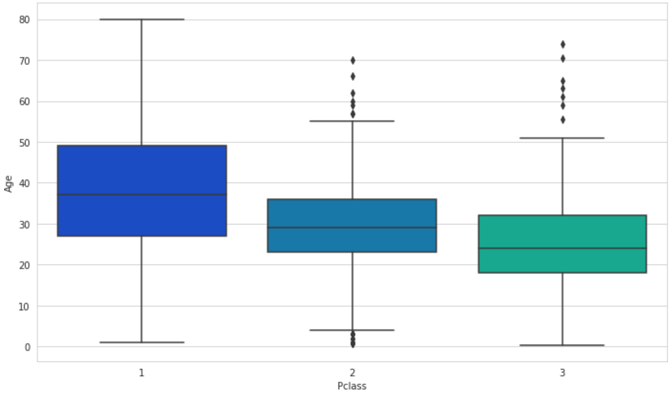
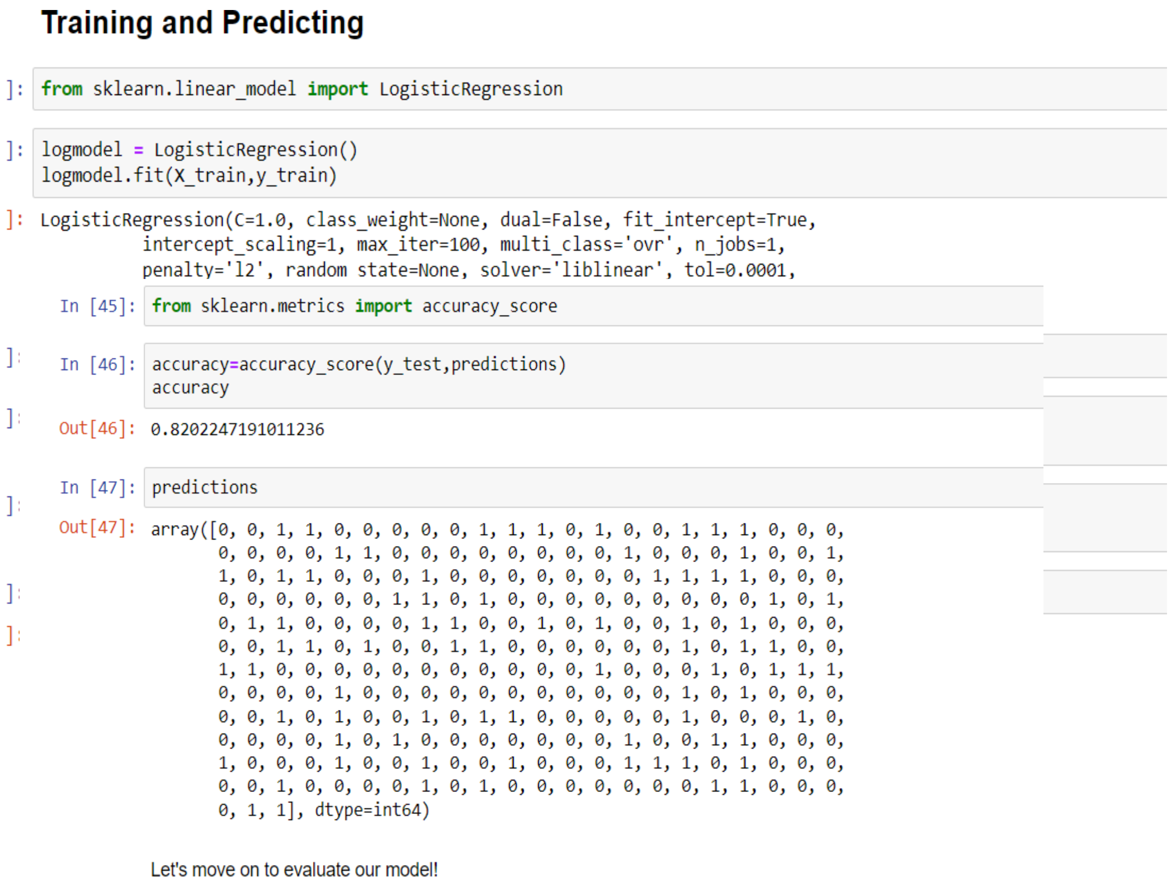


Fig 13 : Boxplot

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%



Training and Predicting

```
]: from sklearn.linear_model import LogisticRegression
```

```
]: logmodel = LogisticRegression()
   logmodel.fit(X_train,y_train)
```

```
]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
             intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
             penalty='l2', random state=None, solver='liblinear', tol=0.0001,
```

```
In [45]: from sklearn.metrics import accuracy_score
```

```
In [46]: accuracy=accuracy_score(y_test,predictions)
         accuracy

Out[46]: 0.8202247191011236
```

```
In [47]: predictions

Out[47]: array([0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0,
                0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,
                1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0,
                0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,
                0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0,
                0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0,
                1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1,
                0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0,
                0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0,
                0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0,
                1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0,
                0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0,
                0, 1, 1], dtype=int64)
```

Let's move on to evaluate our model!

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 0.8272. 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.

REFERENCES

[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