**Titanic Project**

Contents

1. Problem Statement
2. Data Analysis
3. EDA Concluding Remark
4. Preprocessing PipeLine
5. Building Machine Learning Model
6. Concluding Remark
7. **Problem Statement**

   The Titanic Problem is based on the sinking of the ‘Unsinkable’ ship Titanic in early 1912. It gives you information about multiple people like their ages, sexes, sibling counts, embarkment points, and whether or not they survived the disaster. Based on these features, you have to predict if an arbitrary passenger on Titanic would survive the sinking or not.

This is a standard supervised classification problem.

Dataset contains 12 columns(including target variable) and 891 rows.

The target variable (Survived) is in the binary form where 1 is for survived and 0 for non survived.

**Features :**

Passenger Id

Survived

P-class

Name

Sex

Age

SibSp

Parch

Ticket

Fare

Cabin

Embarked

1. **Data Analysis**

Data is from one of the most infamous shipwrecks in history because that two columns cabin and age contains null values.

**2.1** **Data Description**

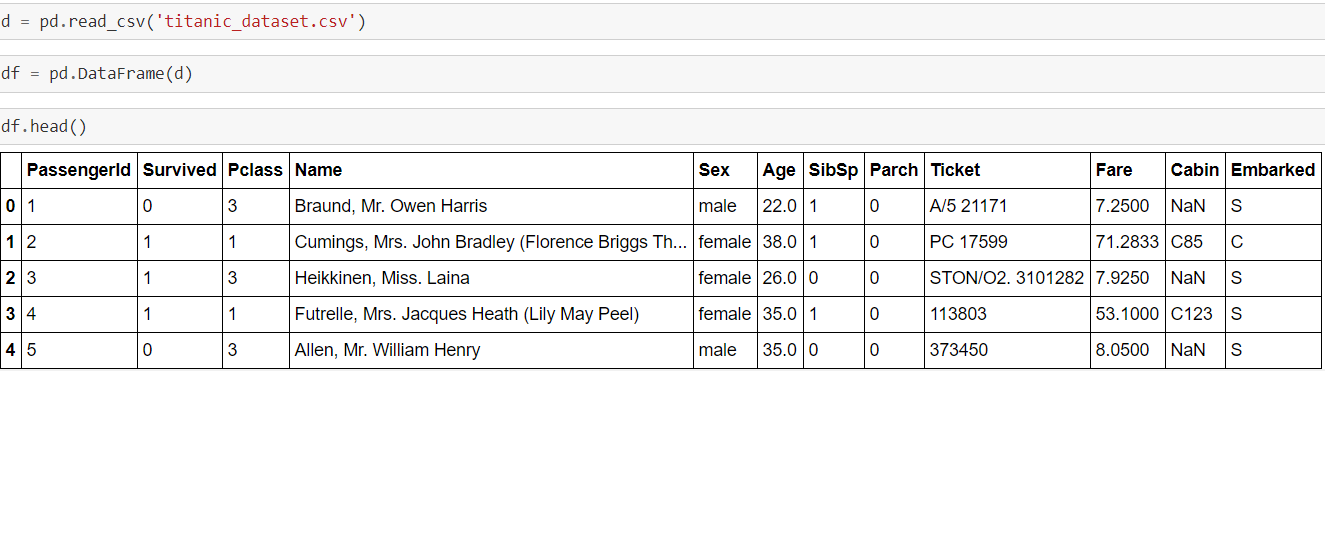
First, the data is imported from the source location and converted into the Data Frame for further analysis.

It is clear from the Head of the Data Frame that SibSp, P-class, Survived,

Id, Age, and Parch are discrete values.

Fare is a continuous column.

Name, Cabin, Embarked, Sex, Ticket is in object Datatype.

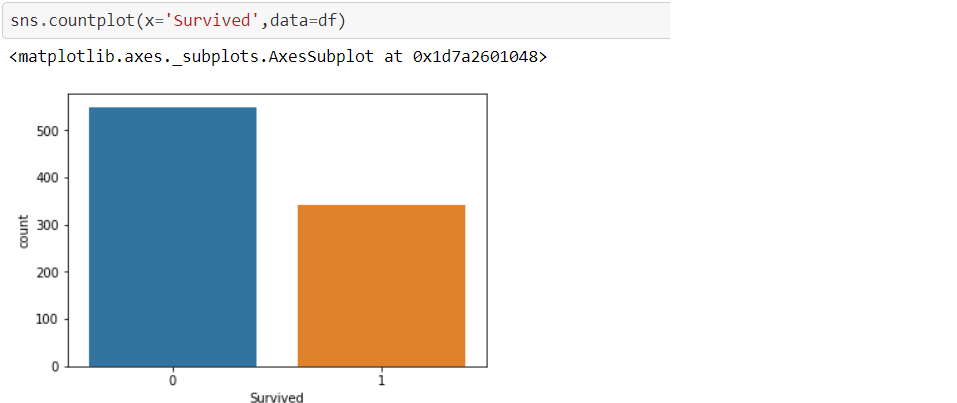


**2.2 Data Visualization**

**2.2.1 Survived**

0 is showing the number of people who not survived and 1 is showing the number of people who survived.

The number of people who survived is less than the number of people who have not survived.



**2.2.2 Survived count based on sex**

The Majority of the people not survived are male.

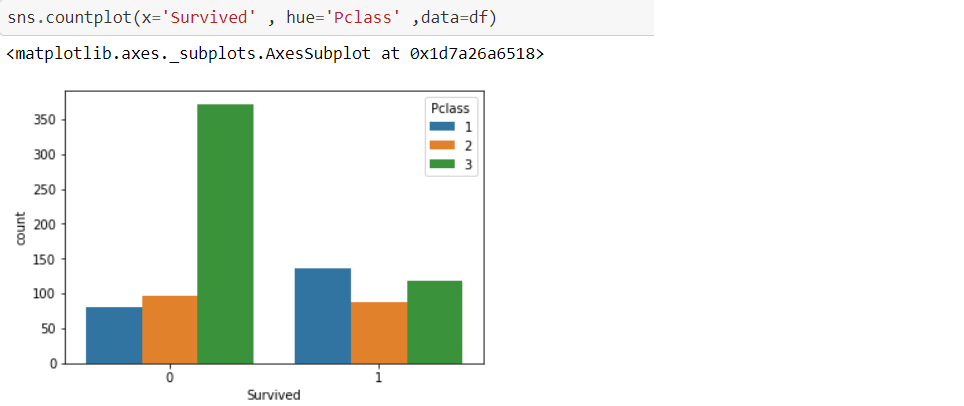
The blue graph is for males and orange for females.



**2.2.3 Survived count based on P-class**

The Majority of the people who not survived are from P-class 3

and Majority of the people who survived are from P-class 1



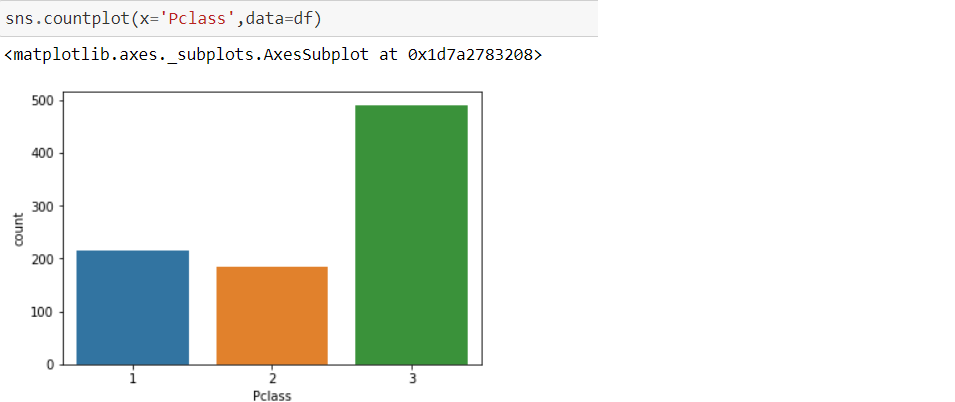
**2.2.4 Survived count based on Age**

People who survived are mostly from 20 to 30 years of age.



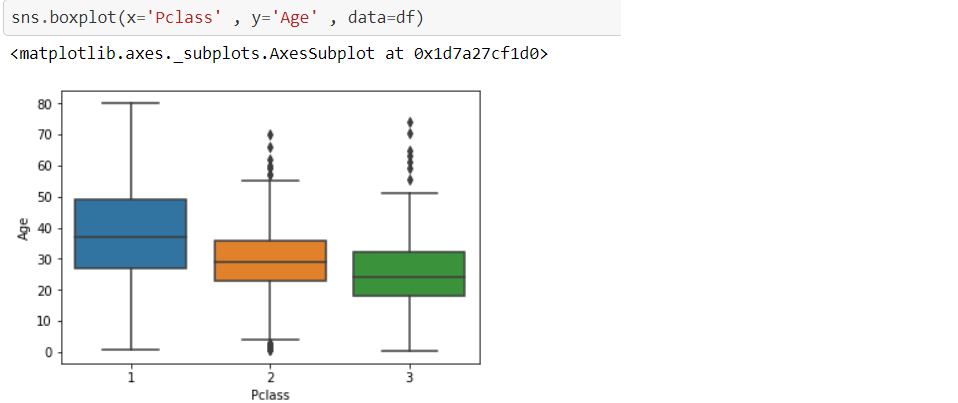
**2.2.5 P-class**

The highest number of people is in P-class 1 and the least number of people in P-class 2.



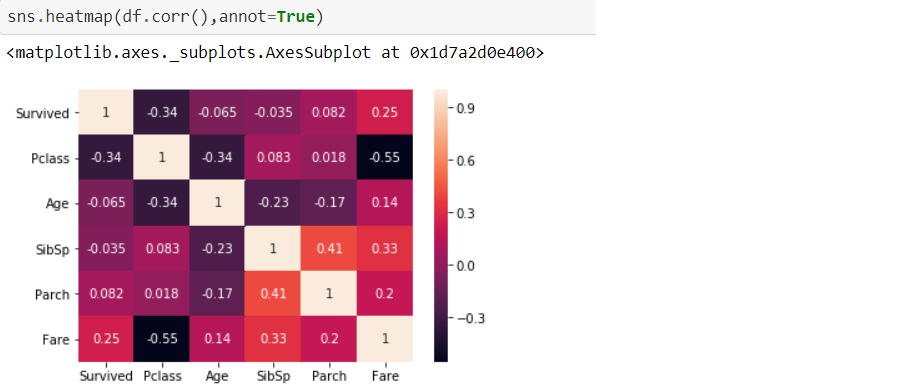
**2.2.6 P-class and Age**

In P-class 2 and P-class 3 people above 60 years are also there



**2.2.7 Correlation**

Target Variable is highly correlated with P-class and Fare.



1. **EDA Concluding Remarks**

* Most of the Females survived the tragedy.
* People from P-class 1 are the majority of the people who survived
* In P-class 1 people were below 60 years of age.
* Target Variable(survived) is positively correlated with fare that concludes people who had given more fare possibility of their survival is more.
* People who survived were more likely to be females, belong to P-class and age could be below 20 and more than 60.
* The Average Age of people in the P-class 1 was 37(app.), In P-class 2 was 30 years(app.) and in P-class 3 24 years(app.)

1. **Pre Processing Pipeline**

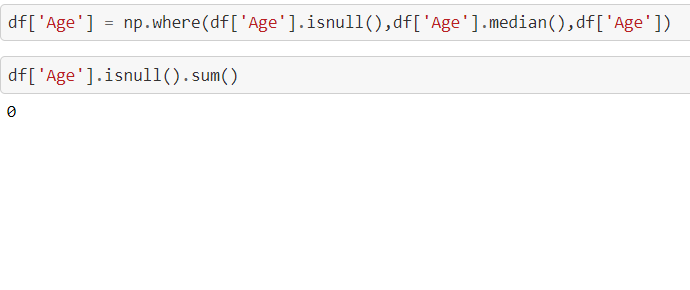
     Pre processing is an essential step to make the dataset ready for the machine learning algorithm.

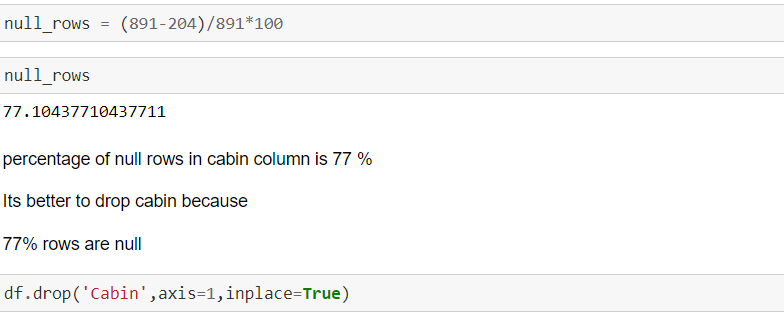
**4.1 Dealing with null values**

Age and Cabin are the columns with null values.

In Age replacing null values with the median.

In Cabin 77% of the values are null so dropping that column.



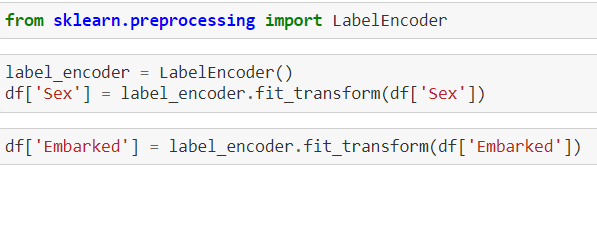


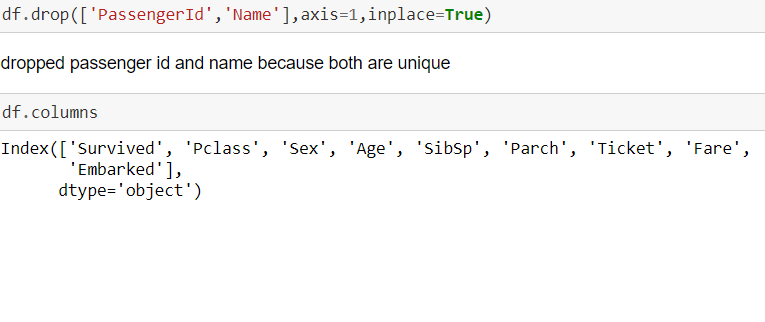
**4.2 Encoding**

Before building a machine learning model it is important for every column to be in numerical form.

Before Encoding I have dropped the name, Id, Ticket column because these columns have only unique values.

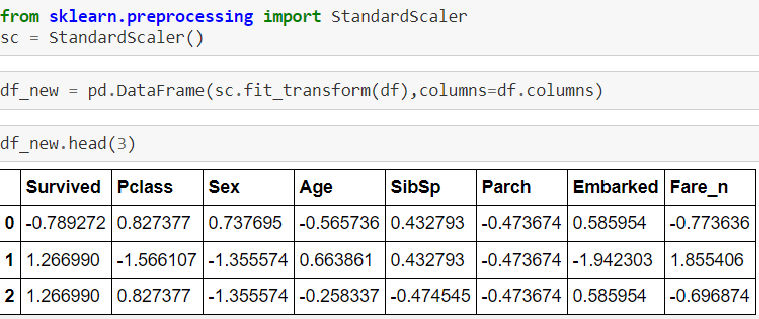
And For Sex and Embarked Column I have used Label Encoder.





**4.3 Scaling**

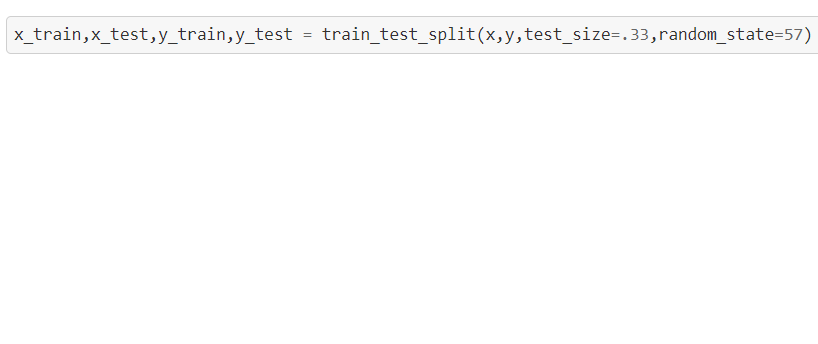
Scaling is important to shrink all the columns in same range for better results.



**4.4 Spitting the Data**

Splitting data into train and test.

Train data is for building the model and Test data for predicting the target variable.



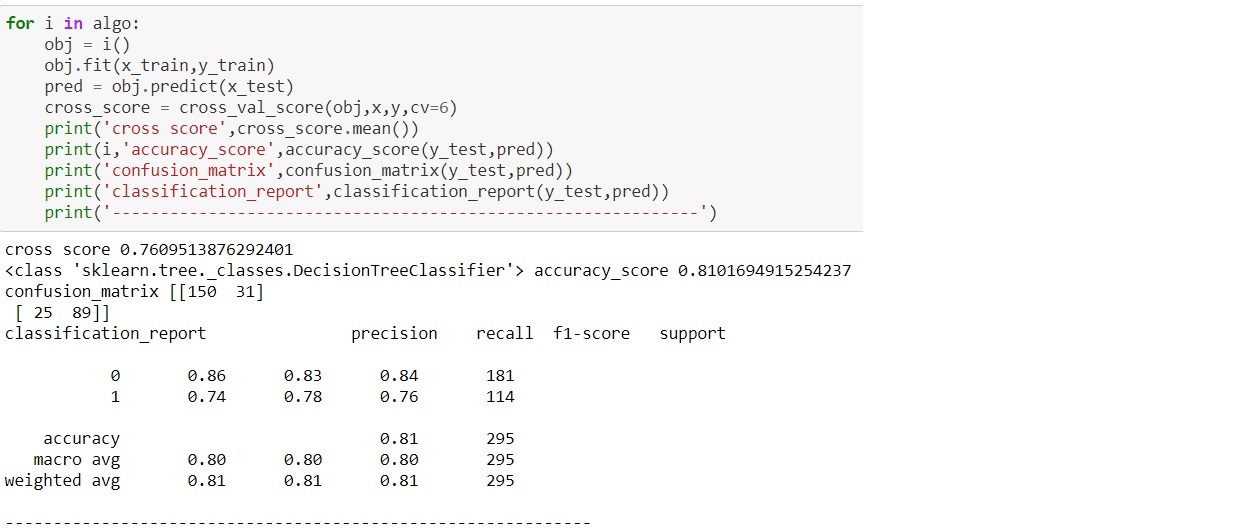
1. **Building Machine Learning Model**

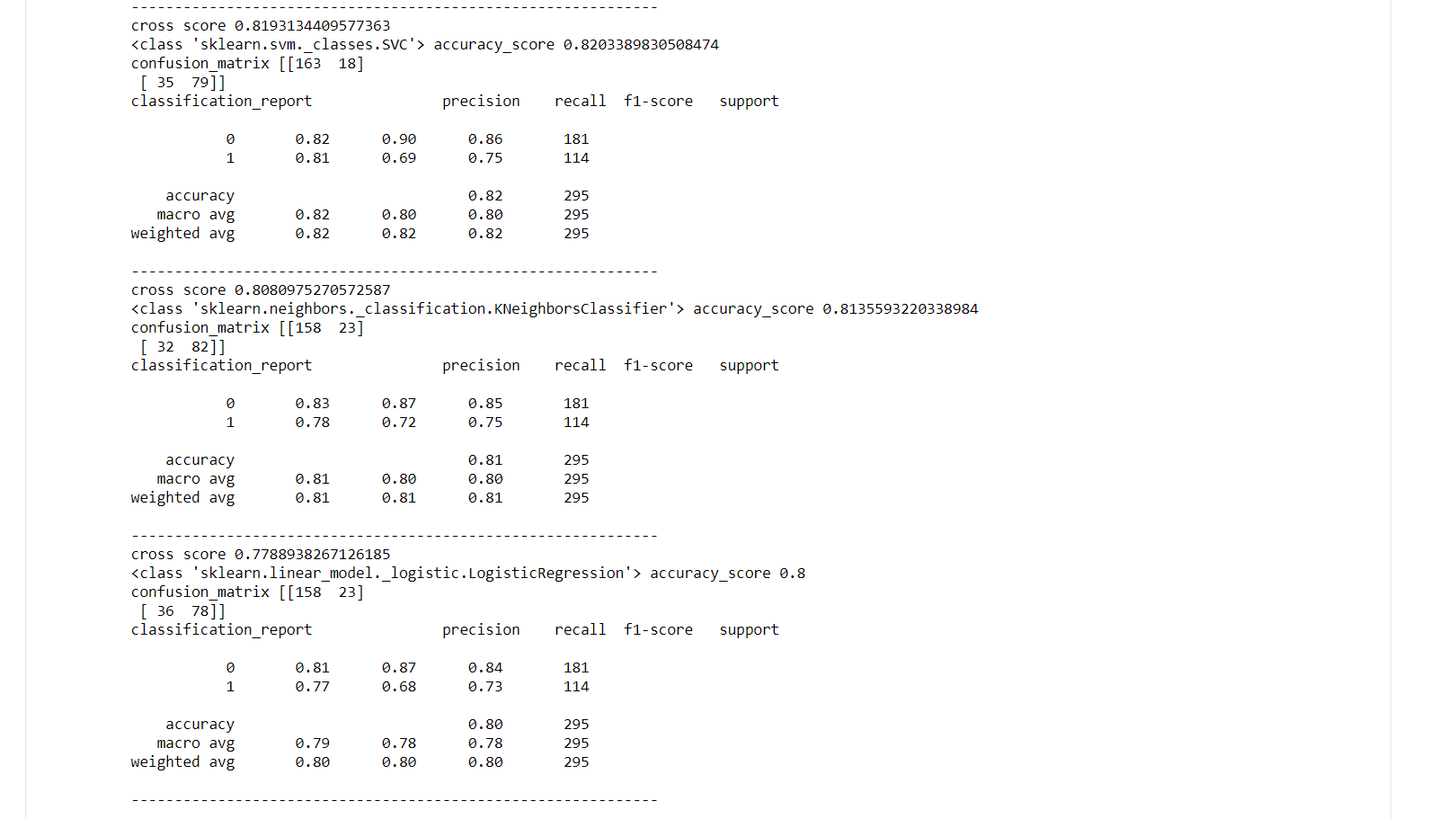
**5.1 Basic algorithms**

I have used some standard classification algorithms like Logistic Regression, Decision Tree Classifier, Kneigbour classifier, and Support Vector Classifier.

I have evaluated these algorithms on the basis of cross-validation score and accuracy score. At least the difference between these two more the model is efficient.

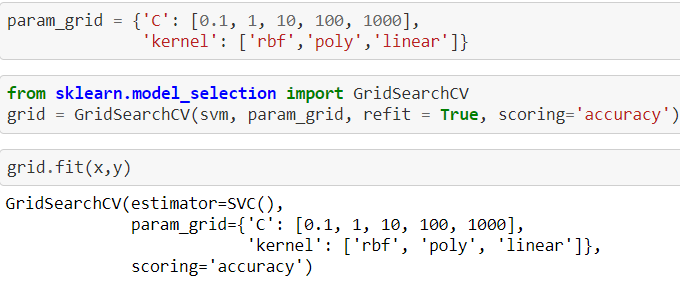
SVC was giving the least difference.





**5.2 Hyper Parameter Tuning**

Moving with the hyperparametric tuning of SVC with GridSeachCV.

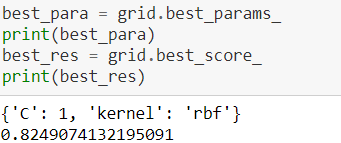


Best Parameters :

'C': 1 and 'kernel': 'rbf'

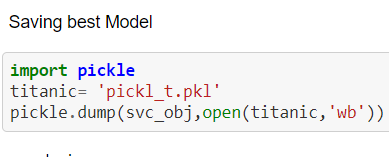
Best Score:

0.8249074132195091



**5.3 Saving the Model**

Saving the model using pickle.



**5.4 Comparing predicted and actual values**



1. **Concluding Remarks**

As the data is from the Titanic sinking incident so there is no scope of adding data for further improvement.

Also, I have deleted four columns because of unique values and the majority of null values to get better results.