Titanic Project Blog

**This article contains the following sub-topics :**

1.Problem Definition

2.Data Analysis

3.EDA Concluding Remarks

4.Pre-processing Pipeline

5.Building Machine Learning Models

6.Concluding Remarks

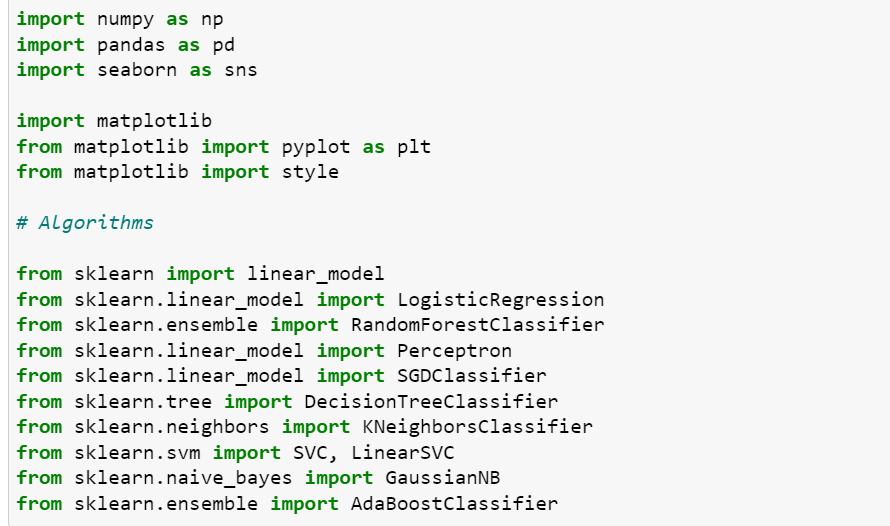
**1.Problem Definition:**

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.

**2.Data Analysis**

**Step 1:**

Import all the necessary libraries and load the datafile into the jupyter notebook. Here we have taken we have loaded the csv files of Titanic and displayed the first few rows so that we have an idea of the columns and data.



* Loading the titanic data set.The dataset contains 891 rows × 12 columns

**Step 2:**

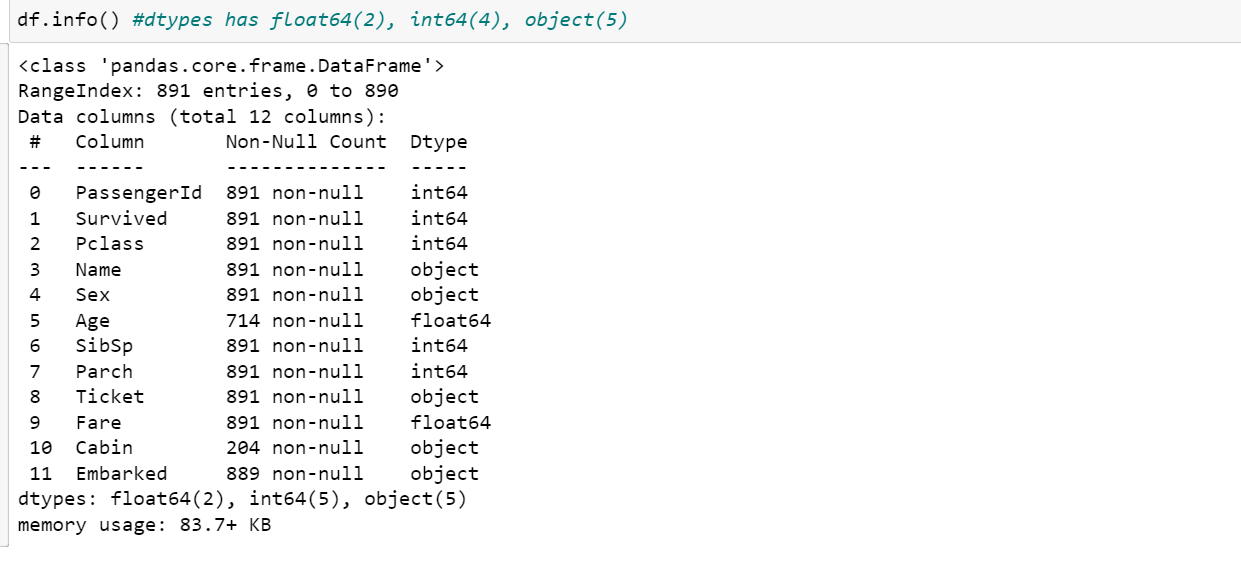
we use df.isnull().sum , df.info() and df.describe() to check if there are any null values in the given datafile and clean the data.

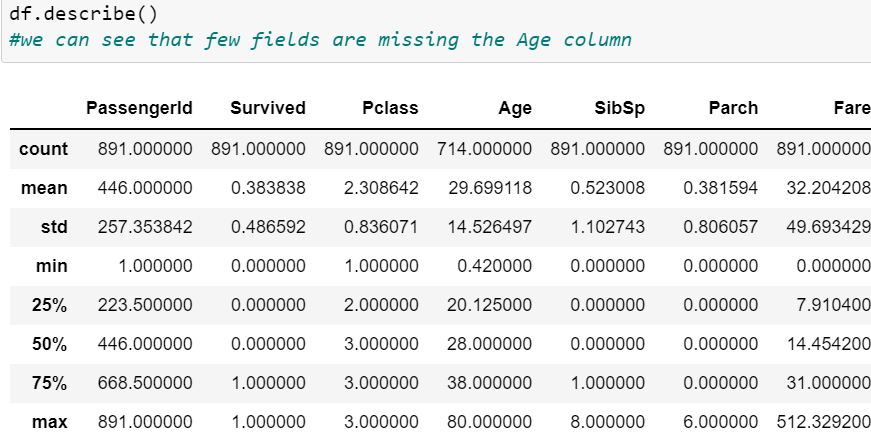
****

**Step 3:**

Let us first look at the columns of the data and then we use describe() and info() methods to get a basic idea of what we have in hand. Here we have different column such as:-**PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,**

**Cabin,Embarked.**

****

****

3.**Exploratory Data Analysis(EDA):**

EDA is an approach of analyzing data sets to summarize their main characteristics, often with visual methods, a statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. we can say that EDA is statistician’s way of storytelling where you explore data, find patterns and tell insights. EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like: - main features of data - variables and relationships that hold between them - identifying which varaibles are important for our problem.

Step 1:

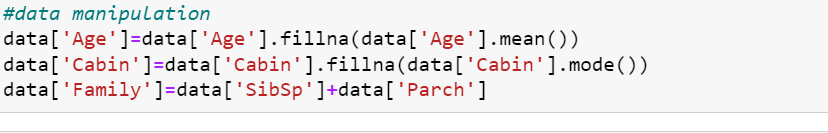
We already know that there are 177 missing values in the age column. From the above results, we see that there are 687 missing values in ‘Cabin’ and 2 missing values in ‘Embarked’. We need to fix these null values before we move on to modelling.



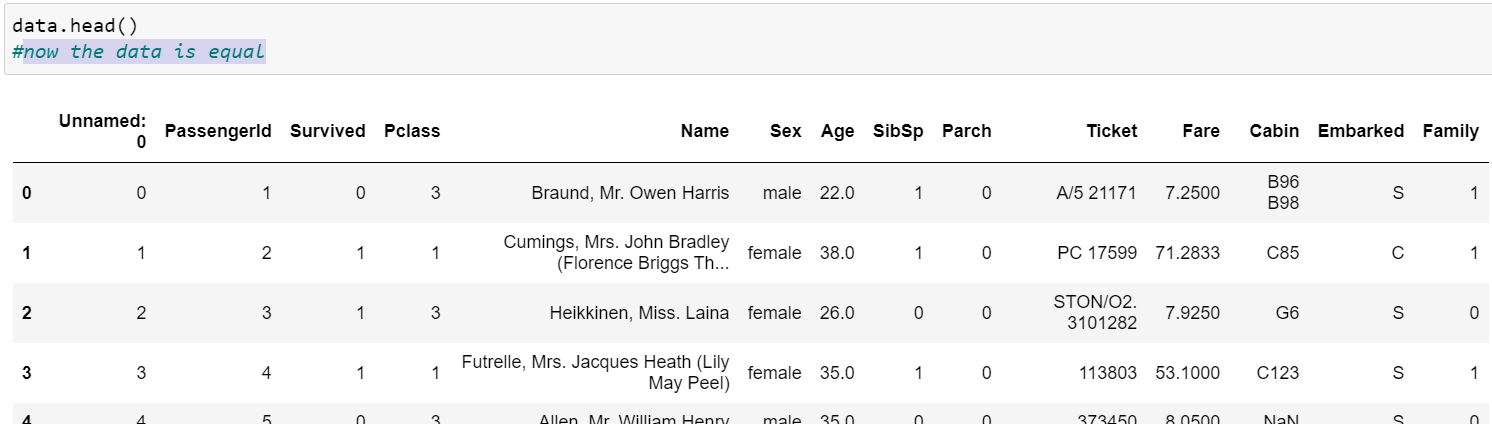
Since the ‘Embarked’ column has only 2 null values, it's not visible in the heat map.

Step 2:

Here we have used fillna() method to take the null values.

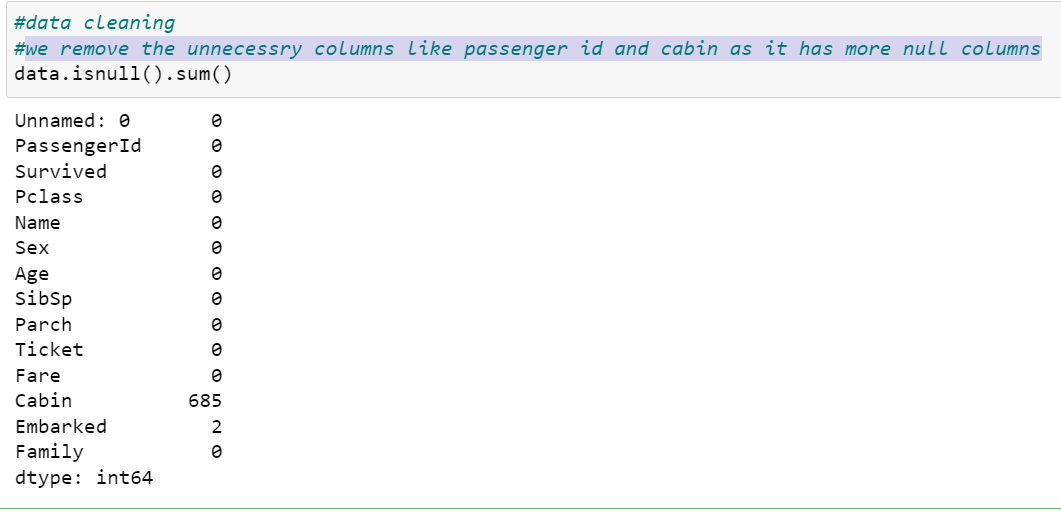


* We can see that now the data is equal.



Step 3:

* **DataCleaning**
  + we remove the unnecessry columns like passenger id and cabin as it has more null columns.

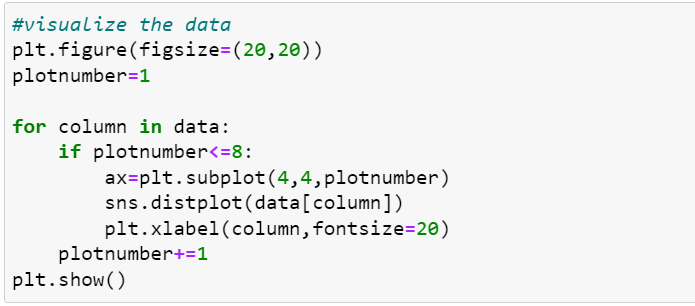


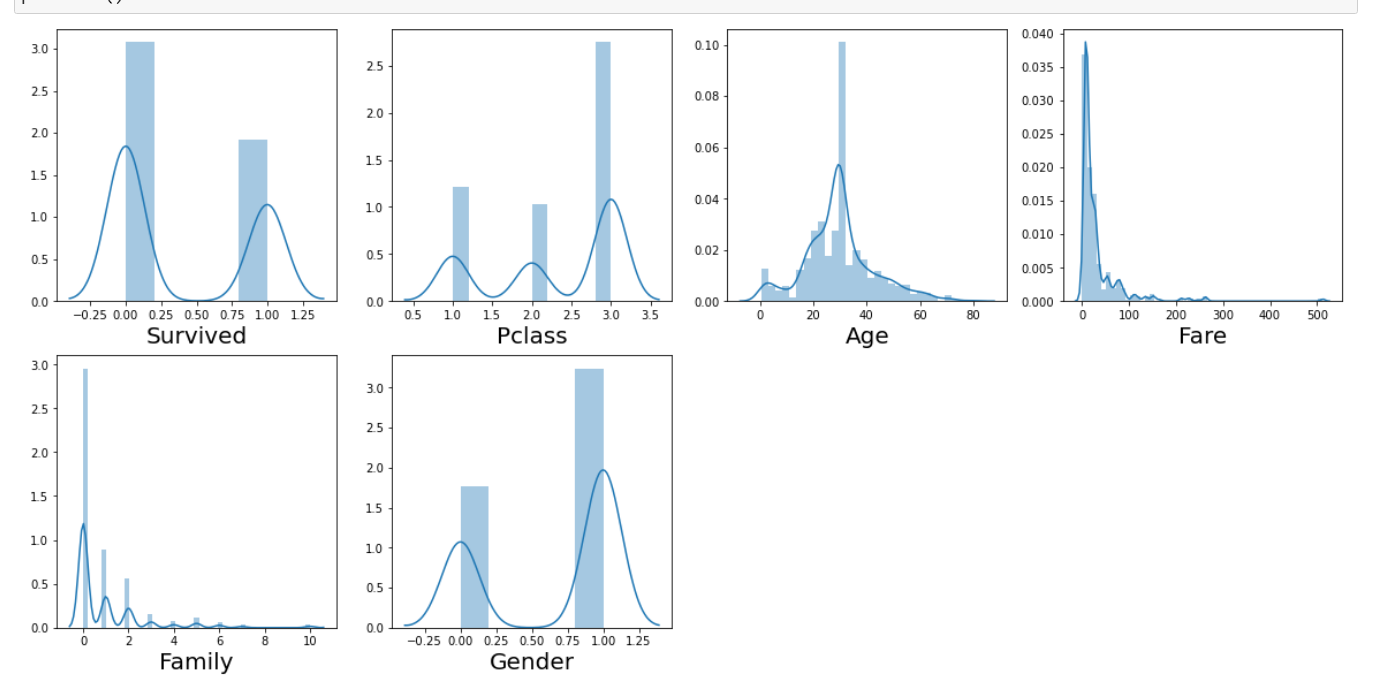


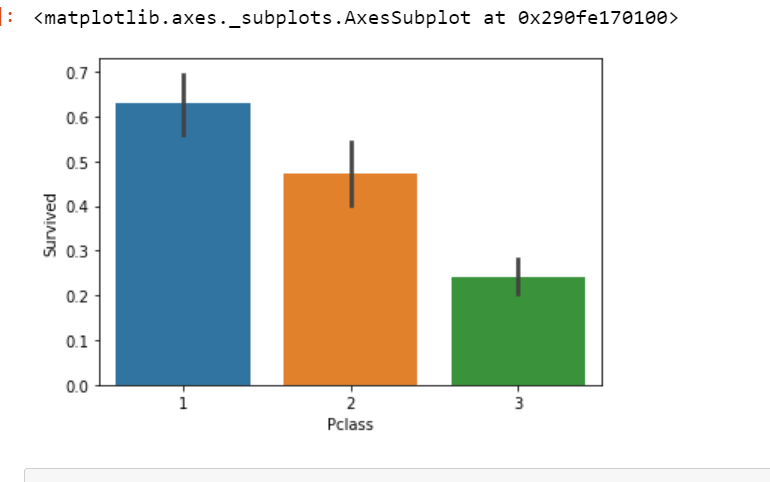
Step 4:

**Data Visualization:**

Visualize the relationship between the label and features, check if it shows positive relation or negative. In this case it is a positive case, so we can go ahead.



* Most of the data is having a normal distribution



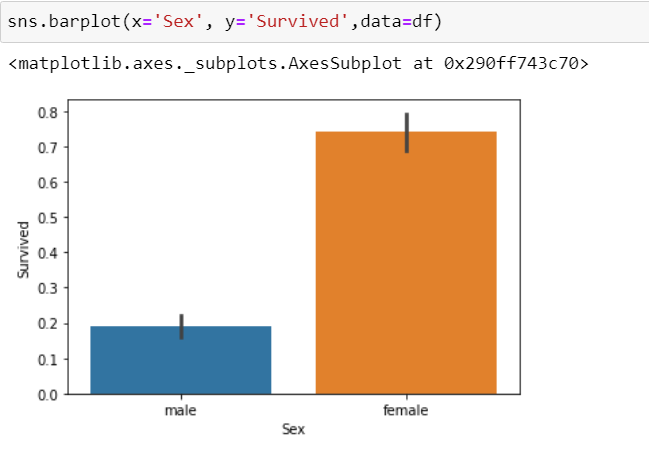
* from this we can observe that no of survivers depend on the class that they are in
* Pclass:3 being the lowest chance or survival

**Pairplot:**

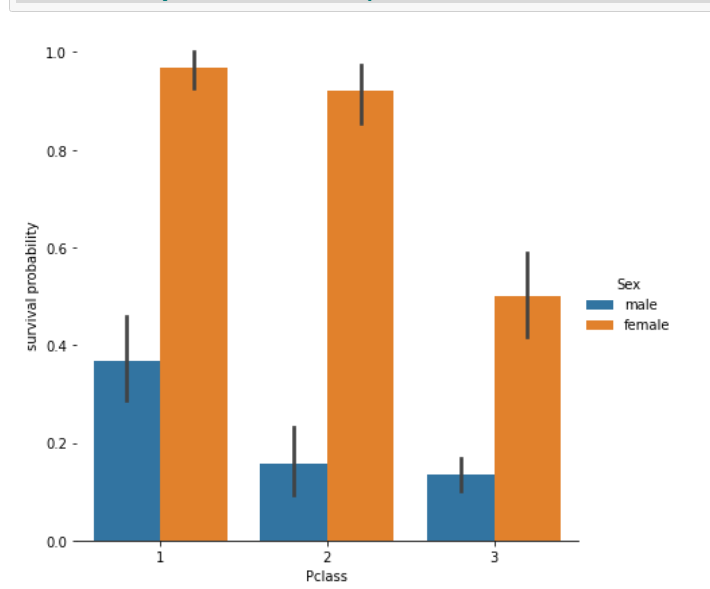


**Barplot:**

* Explored Pclass vs Survived by Sex using Barplot

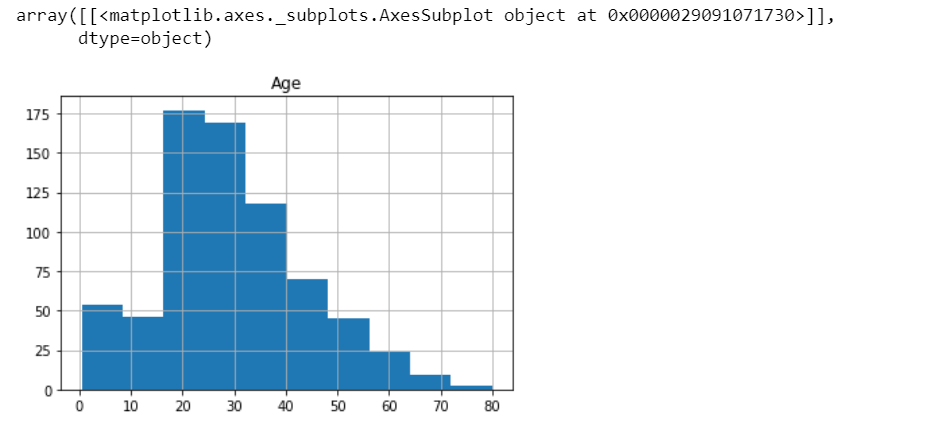
****

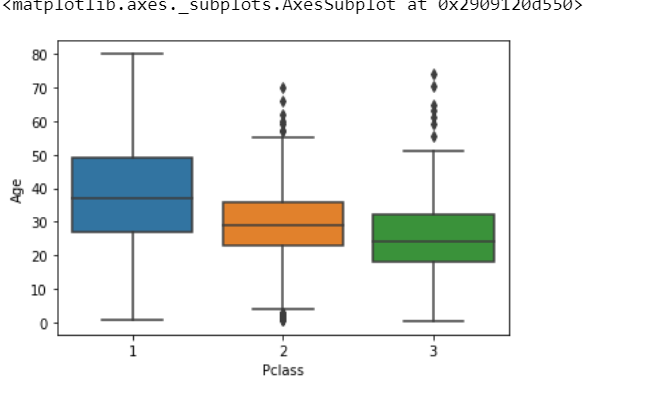
* we can see that there is a high survival rate in female and also the survival rate decreses with Pclass.



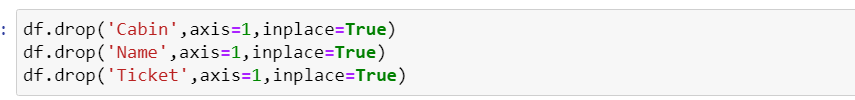
**Histogram:**

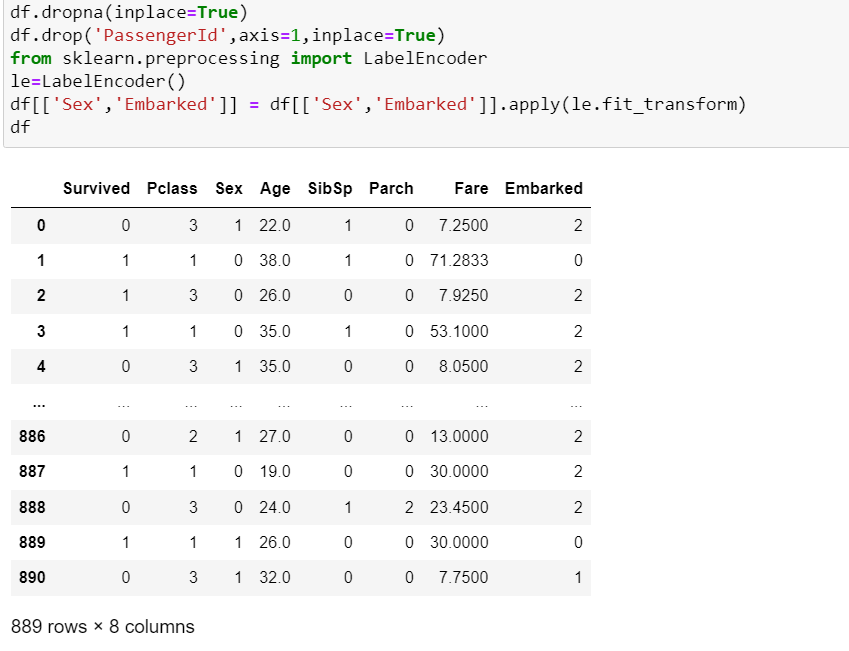
* We can see that most of the survivers are between age 16-48.

* Using **Boxplot** we can observe that as the Pclass incerases the average age also increases



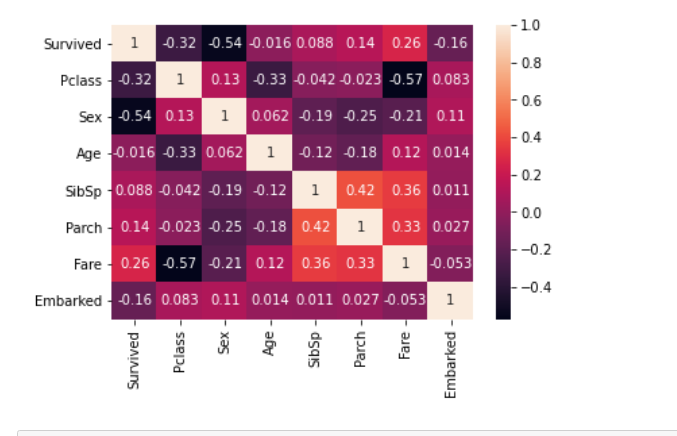
* let us go ahead and drop Cabin column and rows with null values in Embarked





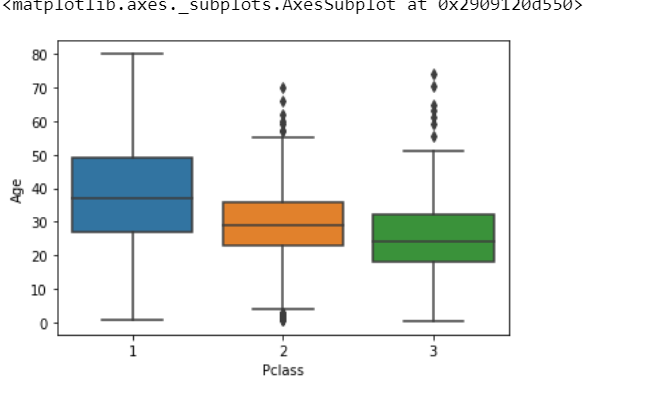
**Correlation:**

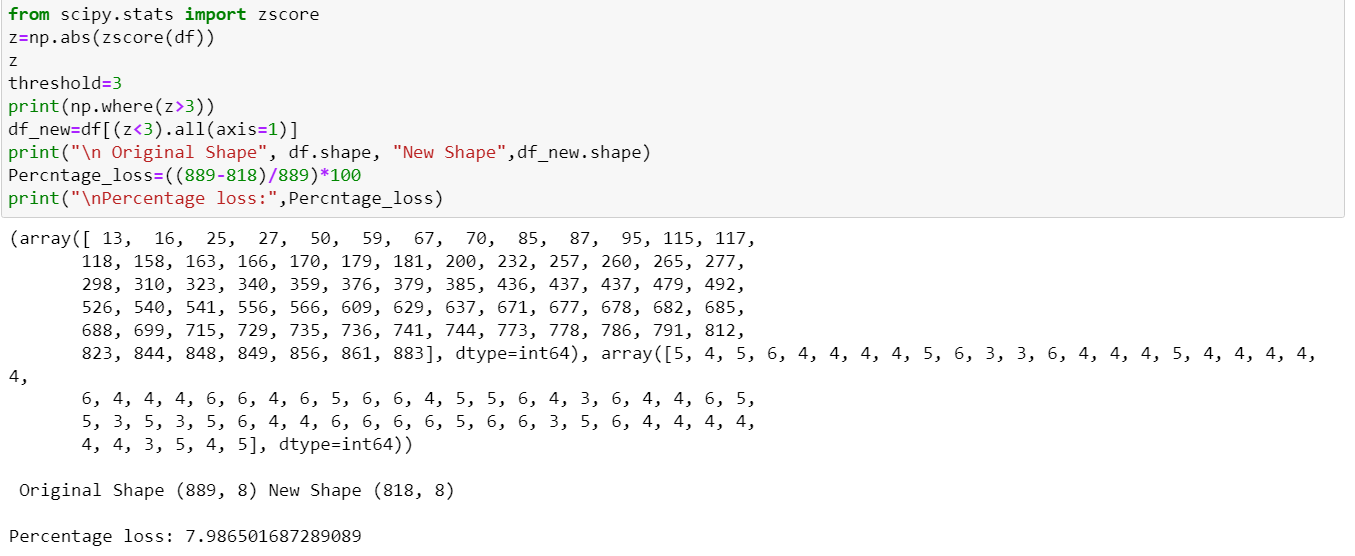
Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables. corr() function is used to see the correlation among the dependent variable and independent variable. You can see correlation in the following figure.



**Outliers :**

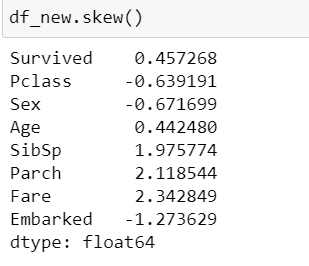
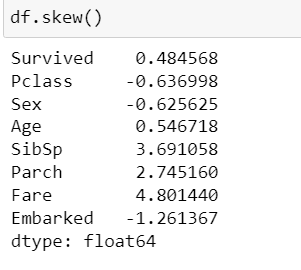
Outliers are the value that lies outside the range of the data, Outliers can be identified through zscore() function from scipy.stats library or sometimes through IQR method, To remove the Outliers we have to select the data that has zscore less than 3 .



* Observing the outliers using Boxplot
* Using zscore checking the percentage of outlier data that is to be removed.
* 7.9% of the total data contains outliers, Loosing 7.9% is not at all affordable so dropping the idea of removing outlier.

**Skewness:**

* The data can be right skewed or left skewed if the median or mean is high and data is highly spread it can be observed through the skew() method, if the skew score is negative and greater than 5 it means data is negatively skewed on left side and if the data is more than +5 it means the data is skewed on right side.



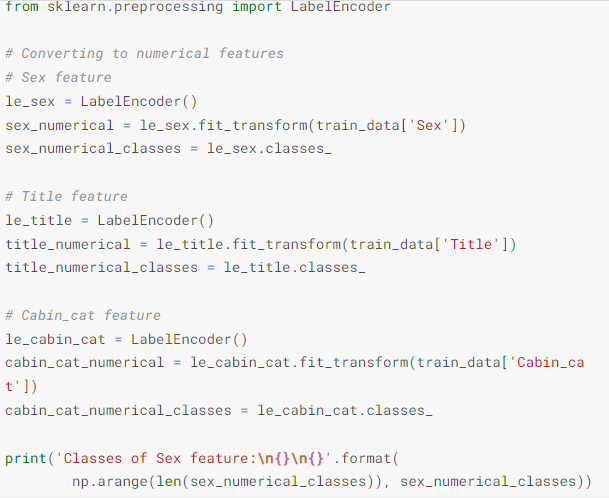
* Here we can see that SibSp,Parch,Fare and Embarked is more than -0.5 to +0.5 It is necessary to cure skewness for this PowerTranform () function is useful but it should be done after separating features and target

**4.Preprocessing Pipeline**

Here we will apply some preprocessing that is needed in particular for various machine learning algorithms to efficiently operate on the data. We will start by encoding our categorical features ('Sex', 'Embarked', 'Title' and 'Cabin\_cat') in a numerical format. The features that we are going to encode might contain missing values/NaNs/Nones. These will have to be imputed.

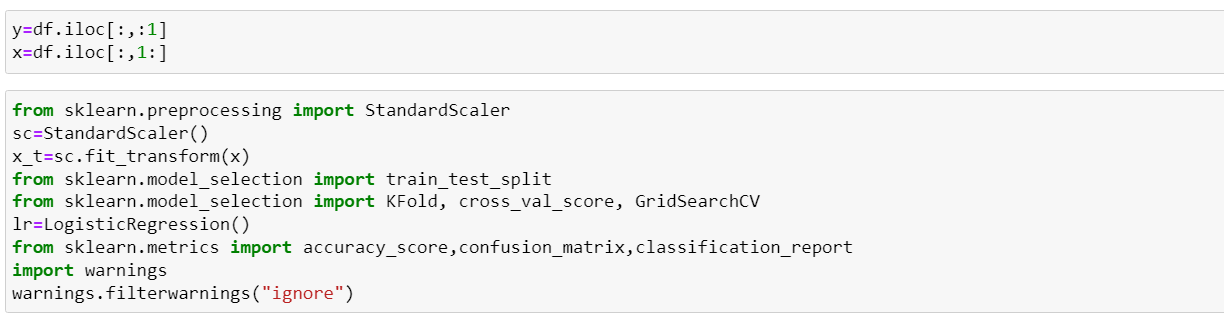
**Encoding:**

We will use sklearn.preprocessing. OneHotEncoder for this; however, OneHotEncoder only works with numerical categorical data. We will use sklearn.preprocessing.LabelEncoder to encode our string labels with numbers.

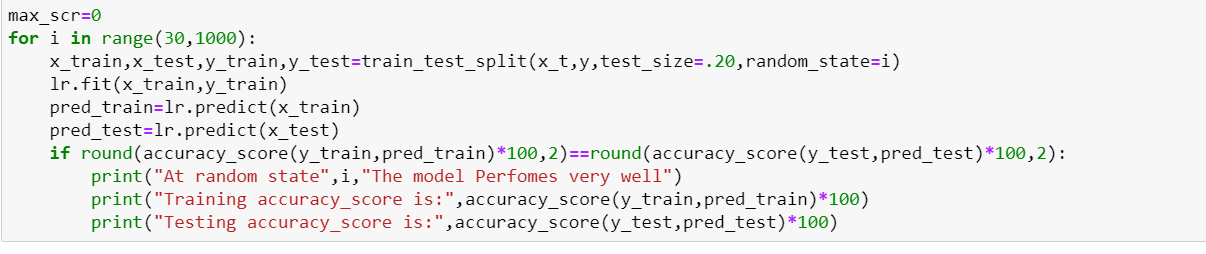


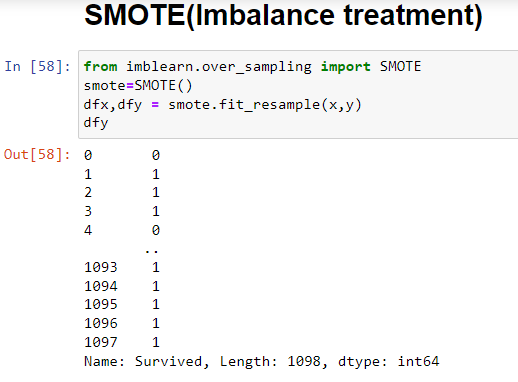
**Feature Scaling**

* Now we can perform standard scaling on all features except the 'Embarked' and 'Age' feature, because both need to be imputed. This scaling will be performed on a temporary copy of the training data because with the sole purpose of being able to more accurately find nearest neighbors for data imputation. Persistent scaling will be performed on the training data in the ML fitting section. Also, the 'Survived' feature does not need scaling since it will be our target label in the ML model training:

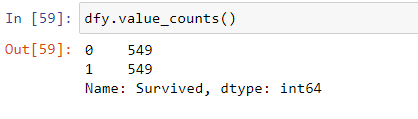


**Splitting data into training and testing sets**

* Split the data into training and test data and apply models and train the data using fit()
* using SMOTE we will balance our target.



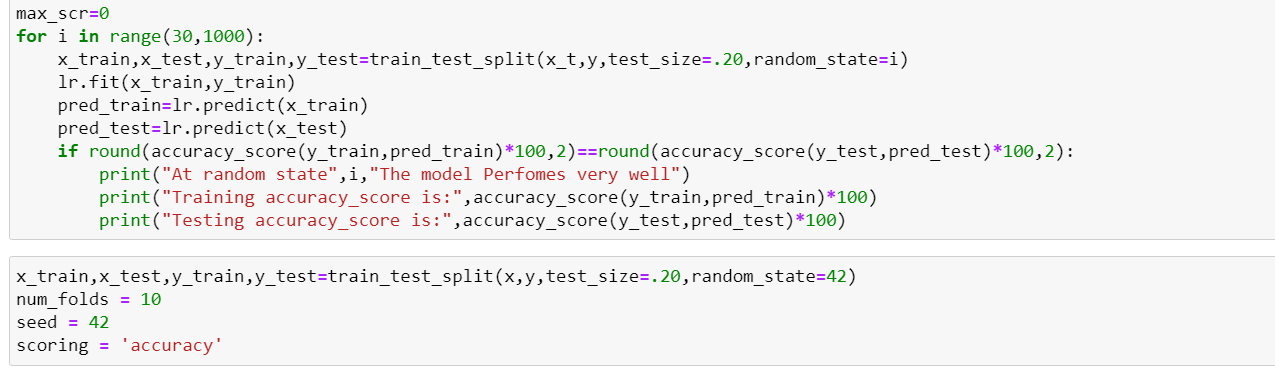
* Here we have used value\_count to see the count of new target variable.Here we can see that its balanced.



**5.Building Machine Learning Models**

**Splitting the Data for Training and Testing**

* Data is splitted into 4 parts for Training and Testing of features ( x ) and for Training and Testing of Target ( y ) like x\_train , x\_test , y\_train , y\_test.

**Training the Models**

To find the best model it is necessary to train different models:

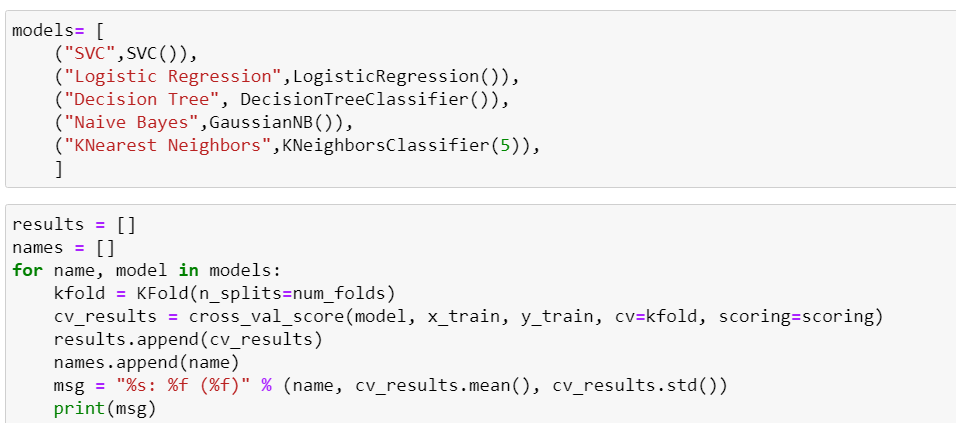
1. Support Vector Machine Classifier

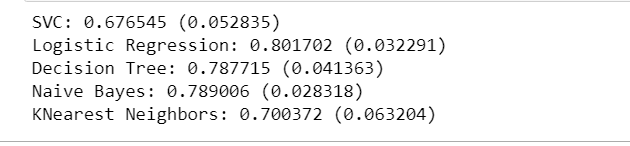
2.Logistic Regression

3. Decision Tree Classifier

4.Naive Bayes Classification

5. KNeighbors Clasifier

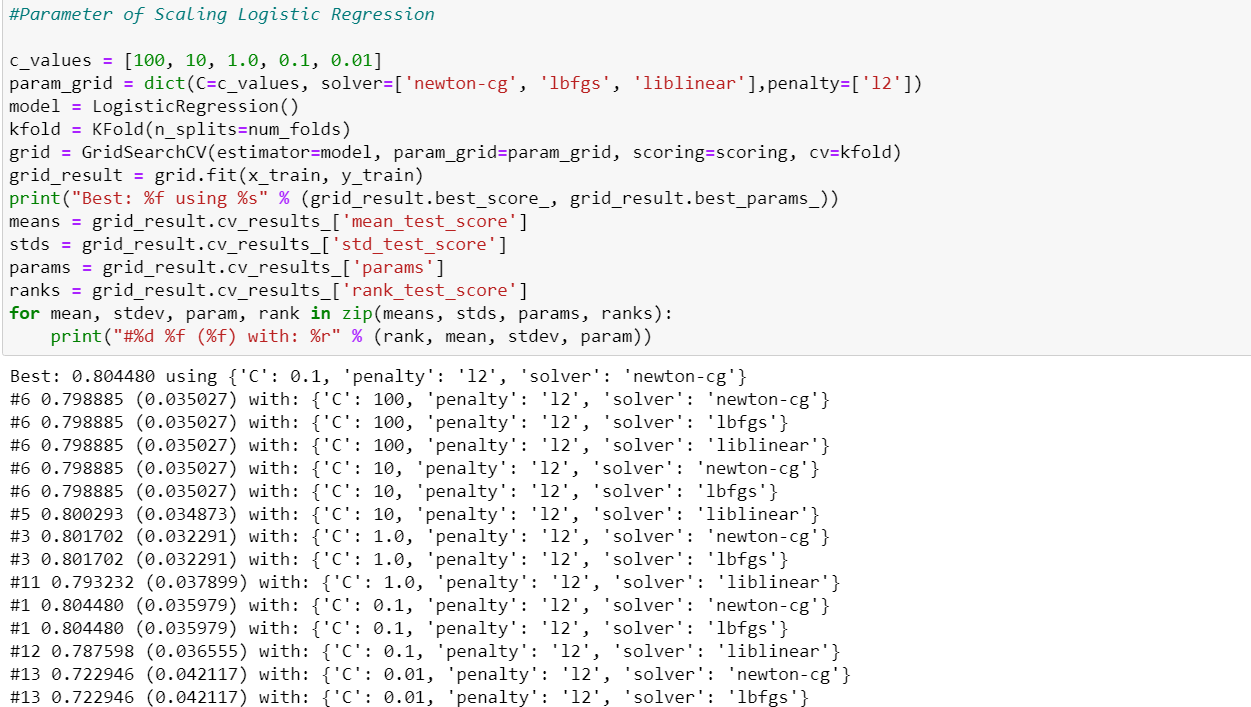




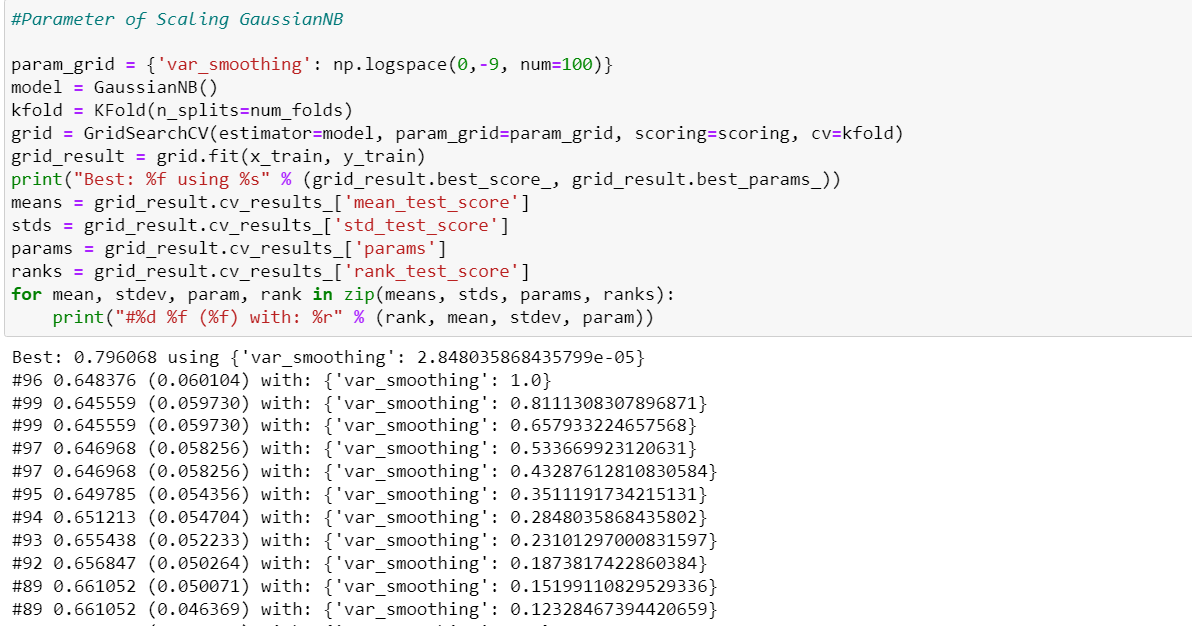
* Logistic Regression: 0.801702 (0.032291) Naive Bayes: 0.789006 (0.028318) are the models performing best with the data

**Scaling the Data:**

* Scaling the parameters Logistic Regression.

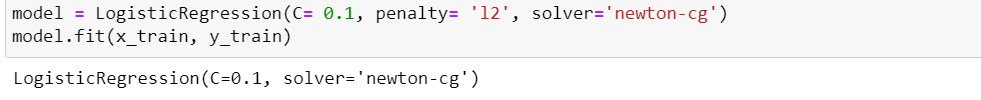
****

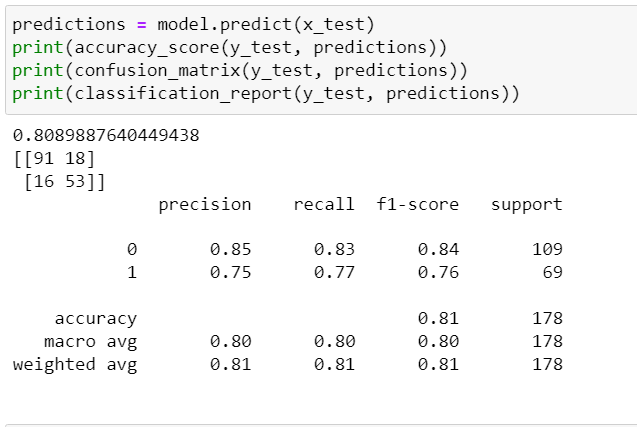
* Scaling parameters of GaussianNB



* we have been able to achive a Best: 80.44% using {'C': 0.1, 'penalty': 'l2', 'solver': 'newton-cg'} As the Parameters in **LogisticRegression.**

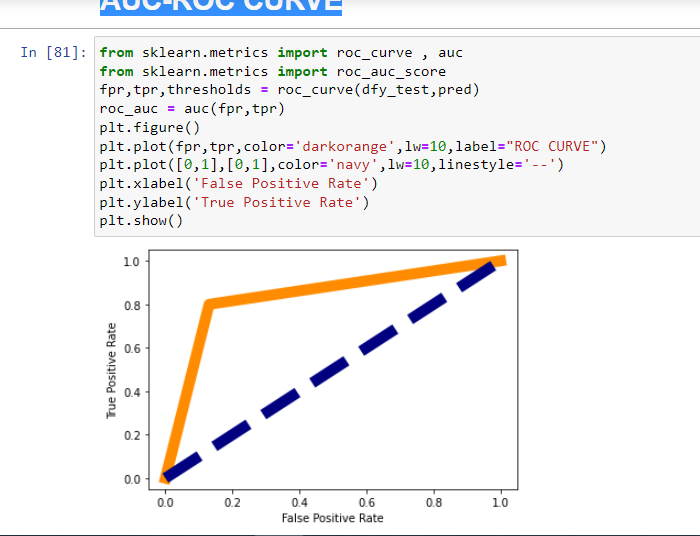
**Training the model using Logistic Regression:**

* Now the model is trained, so test the model by predicting the results for new set of data and compare the actual results and predicted results



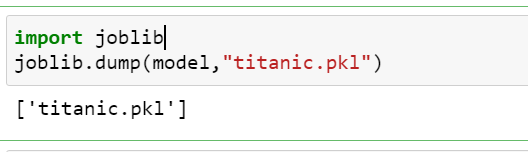
* We can observe that our Model has achieved 80% accuracy.

**AUC-ROC curve:**

****

**Saving the Model:**

* Saving the model into a pickle file, so that we can deploy the model.

****

**6.Concluding Remarks:**

* This project aims to find factors that may affect survival probability of individual passengers and crew when disaster happens. I used 892 training data from Titanic case, while Support Vector Machine Classifier, Logistic Regression, DecisionTree Classifier, Naive Bayes Classification and KNeighbors Classifiers are implemented in this project.
* Through this test, it is obvious to conclude that Logistic Regression can outperform among these three methods, according to the number of mislabeled test points. Also in terms of each factor, sex and passenger class are the two major factors that may determine survival of individual passenger, e.g. female and higher class passengers will be more likely to survive in this case.