**Titanic Project**



# Contents

1. [**Problem Definition**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#c099)
2. [**Data Analysis**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#6270)
3. [**EDA Concluding Remarks**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#030a)
4. [**Pre-processing Pipeline**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#dd40)
5. [**Building Machine Learning Models**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#8341)
6. [**Concluding Remarks**](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#13e2)

1.**Problem Definition:**

Titanicwas a British passenger liner operated by the white star line that sank in the North Atlantic Ocean on 15 April 1912, after striking an iceberg during her maiden voyage from Southampton to New York City. Of the estimated 2,224 passengers and crew aboard, more than 1500 died, making the sinking at the time one of the deadliest of the sinking ship and the deadliest peacetime sinking of a super-liner or cruise ship to date. With much public attention, the disaster has since been the material of many artistic works and a founding material of the disaster film genre. It gives you information about multiple people like their ages, sexes, sibling counts, embankments points, and whether or not they survived the disaster.

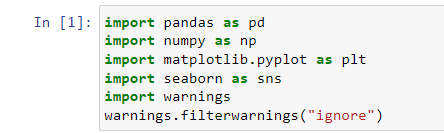
**Analytics** or **People Analytics**-

In this study, we will attempt to solve the following problem statement is:

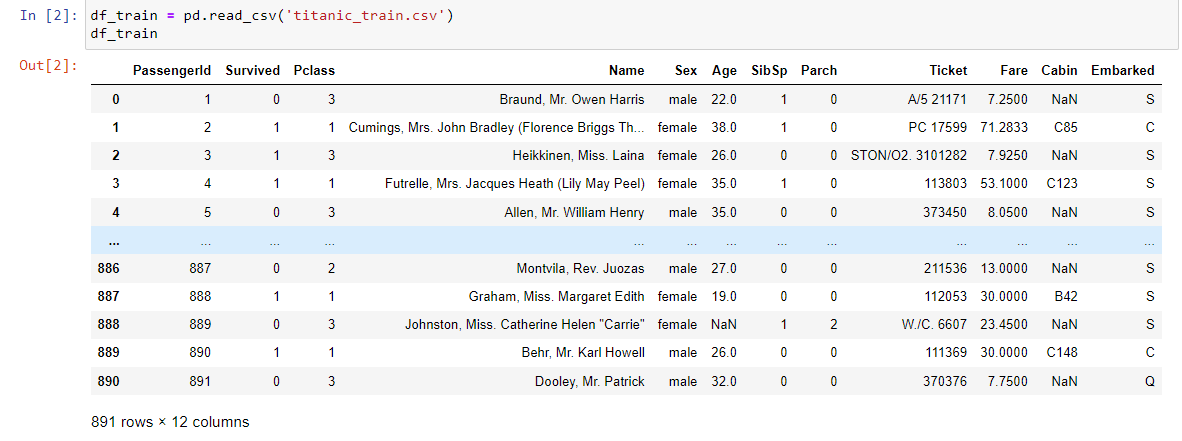
* How many people were saved?
* Which class people were survived, which class they belong.

**2.Data Analysis**

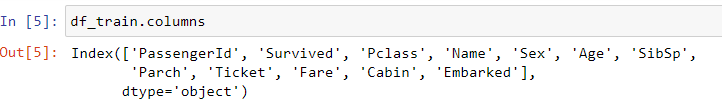
First, we will import the necessary packages and load the data set. The dataset contains 891 rows × 12 columns

****

Here we have loaded the csv files of Titanic

****

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,

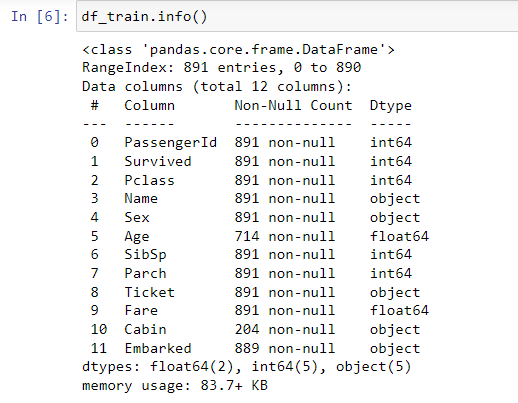
Cabin,Embarked. ****

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

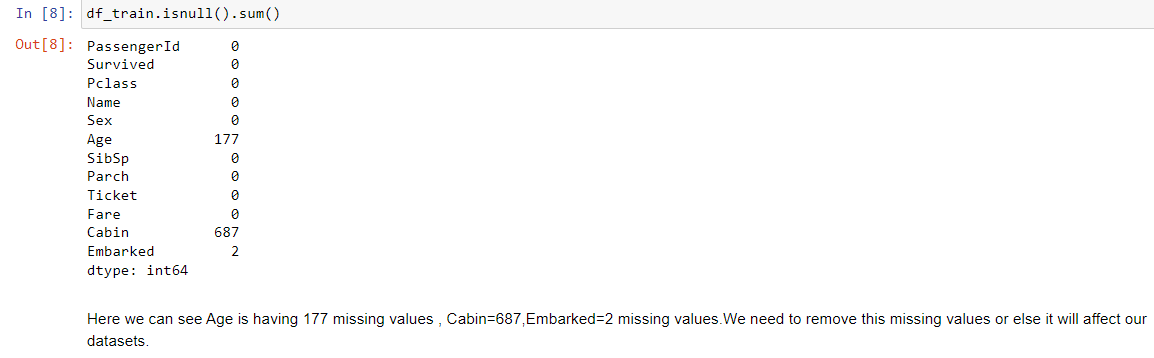
EDA is an approach of analyzing data sets to summarize their main characteristics, often with visual methods, a statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. we can say that EDA is statistician’s 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

we shall look at various exploratory data analysis methods like –

**3.1. Handling/ Removing null values/ missing data –**

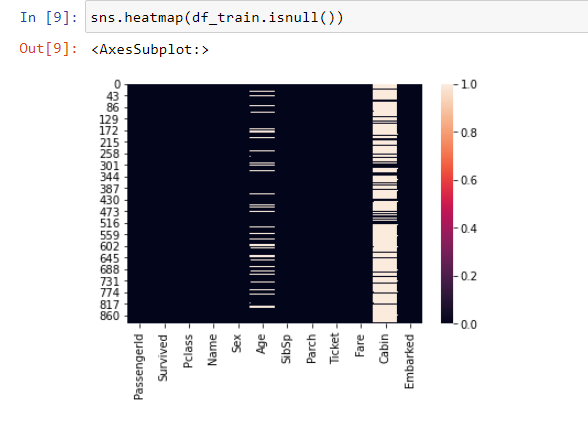


Using the info() method, we get the dtypes of the dataset, Non-null count, and memory usage. To get the total null values of the data we have to use isnull() method.

****

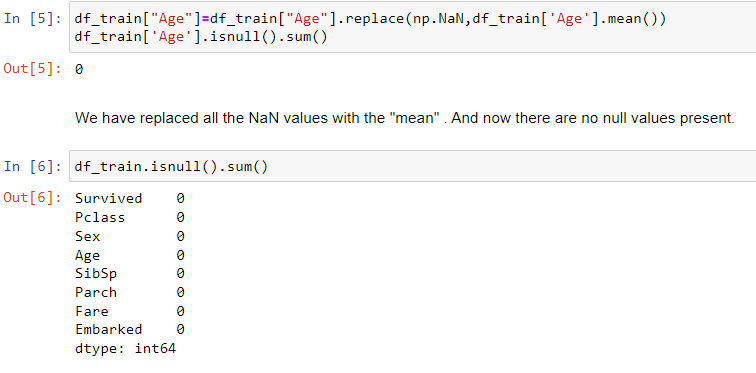
We already know that there are 177 missing values in the age column. From the above results, we see that there are 687 missing values in ‘Cabin’ and 2 missing values in ‘Embarked’. We need to fix these null values before we move on to modeling.

Now let's see the heat map of the null values.

****

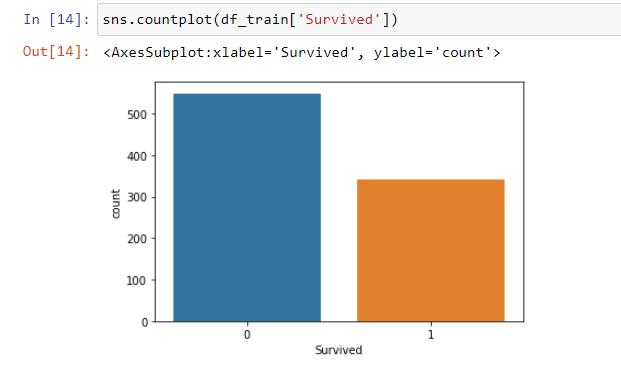
Since the ‘Embarked’ column has only 2 null values, it's not visible in the heat map.

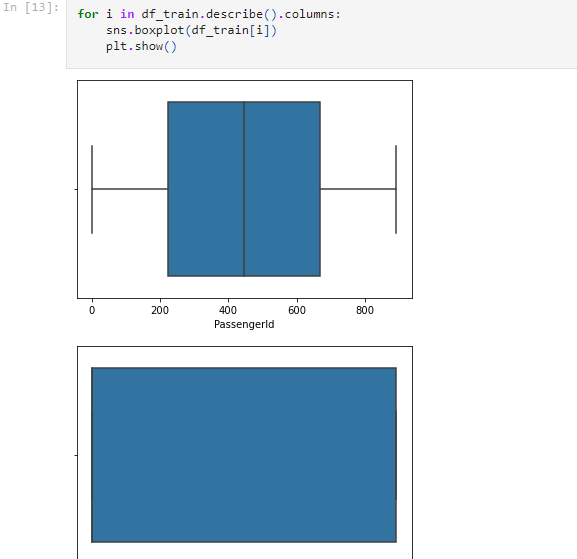
Here we have used replace method to take the null values



**3.2. Visualization of dataset:**

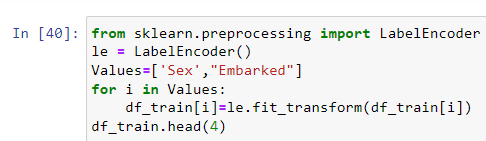
It consist of exploring the dataset visually through histogram, bar chart usally for single variables, Like its better to use countplot for the categorical value and histogram or distplot for continuous variable like for all the categorical values like Name,Sex,Ticket,Cabin,Embarked. Count plot we can use Survived,PassengerId,Pclass,Age,Sibsb,Parch,Fare etc.

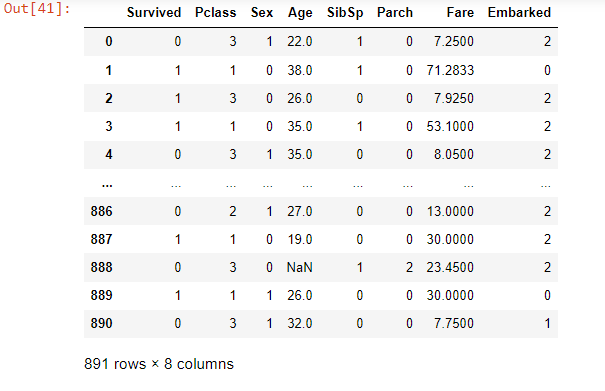




**3.3. Handling Categorical variables**

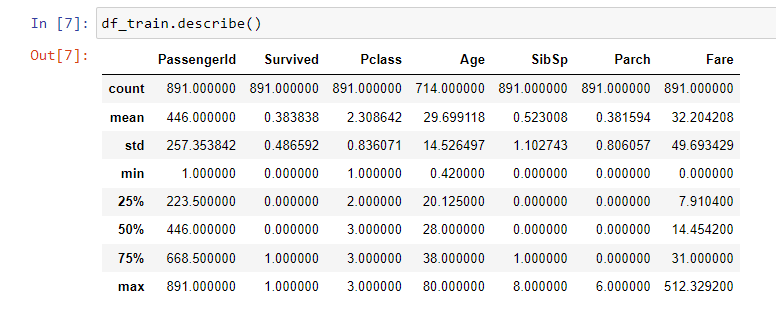
Most statistical model cannot take Objects / Strings as input they only takes numbers as inputs , with LabelEncoder() it is possible to categorize the string into Numbers as 1,2,3 and so on like Name,Sex,Ticket,Cabin,Embarked have string input they all are converted through LabelEncoder().

****

****

* 1. **Descriptive Statistics** –

It is a way of giving a brief overview of the dataset we are dealing with , including some measures and features of the sample.   
 For this we can use the describe() function the describe() function tells the mean() , min(), max(), 25th percentile , median 50th percentile and 75th percentile of the data.



Let us look at the data row-wise and try to understand.

**Count:** The first-row ‘count’ is the number of entries in that particular column ‘PassengerID’. It shows 891(equal to the number of passengers), which means ‘PassengerID’ is available for all the passengers. Except ‘Age’, all other data are available for each passenger. We gotta do something to fix the missing values in the ‘Age’ column.

**Mean:** Mean of all data. Let's consider the ‘Fare’ column. The average ‘Fare’ of the passenger is USD 32.20

**Std:** Standard deviation of all data. A low standard deviation means that most of the numbers are close to the average, while a high standard deviation means that the numbers are more spread out. Again ‘Fare’ has a high standard deviation.

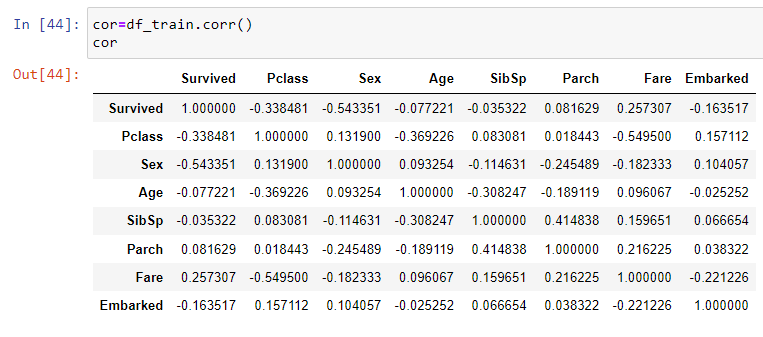
**Min:** Minimum value of the column. For example, the lowest ‘Fare’ shows USD 0, which means ‘Fare’ is unavailable for some passengers. The model may not perform well with 0 for some passengers. Hence we will need to focus on the ‘Fare’ column before modeling.

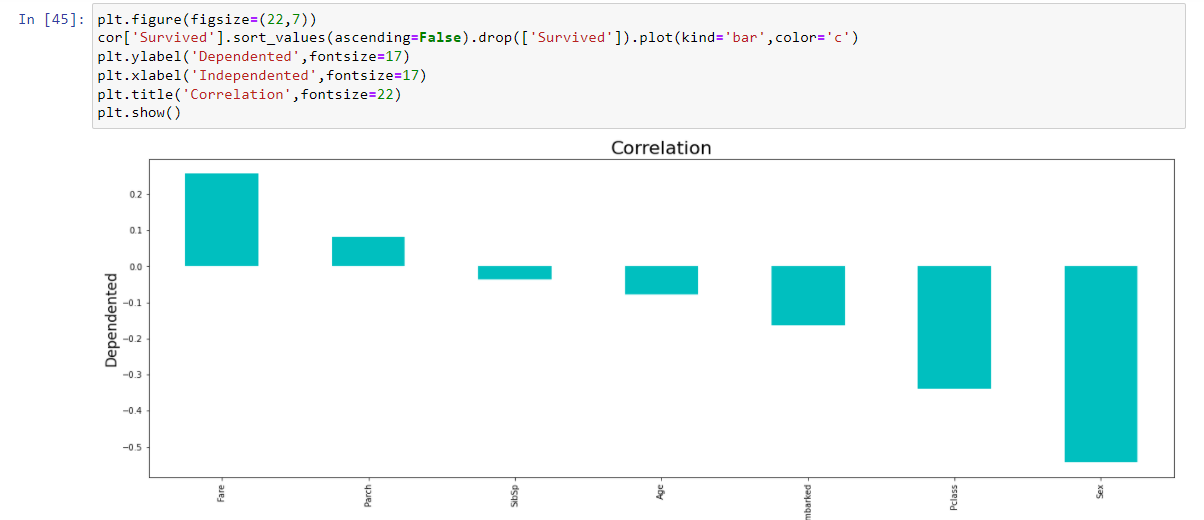
**Max:** Maximum value of the column. For example, the highest ‘Fare’ shows USD 512.33. The mean of the ‘Fare’ column says USD 32.20. We see there is a huge difference in fare. It could be because of the ‘Pclass’ they were traveling.

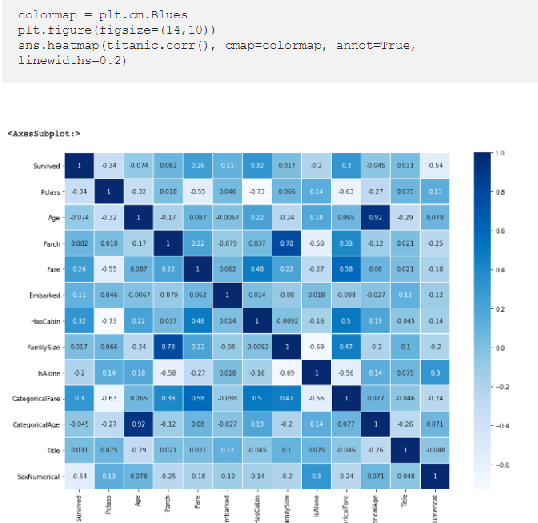
**25%, 50% & 75%:**1st, 2nd, and 3rd quartile of the data. Quartile in statistics is a type of quantile that divides the number of ordered data points into four equal groups. 1st quartile is the middle number between the smallest number and the median. 2nd quartile is the median of the dataset. 3rd quartile is the middle value between the median and the highest value.

* 1. **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



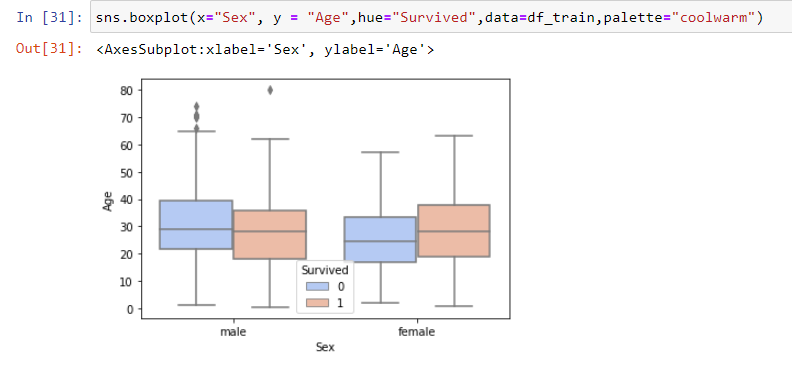


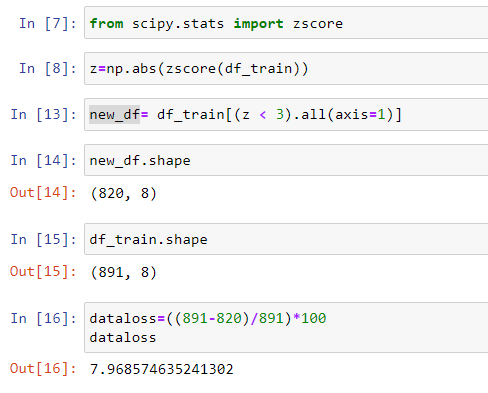


The first row contains the values that represent the correlation of each variable with the target variable. ‘HasCabin’ and ‘CategoricalFare’ are highly (positively) correlated with the target variable and ‘Sex Numerical’ is inversely correlated with the target variable.

**3.6. 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 .

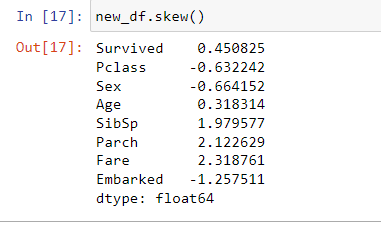




7% of the total data contains outliers, Loosing 7% is not at all affordable so dropping the idea of removing outlier.

* 1. **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.



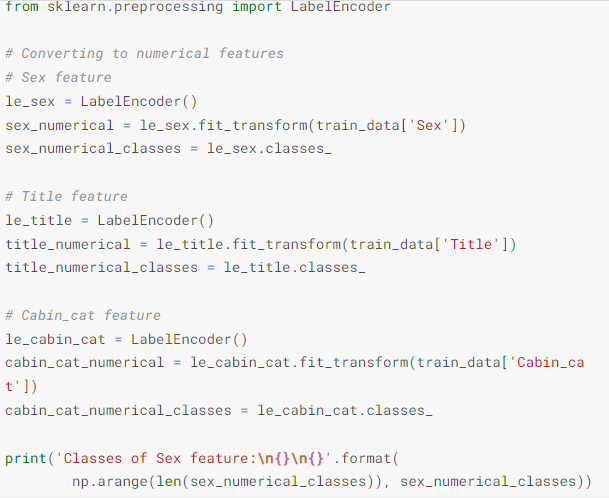
Here we can see that SibSp,Parch,Fare and Embarked is more than -0.5 to +0.5 It is necessary to cure skewness for this PowerTranform () function is useful but it should be done after separating features and target

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.

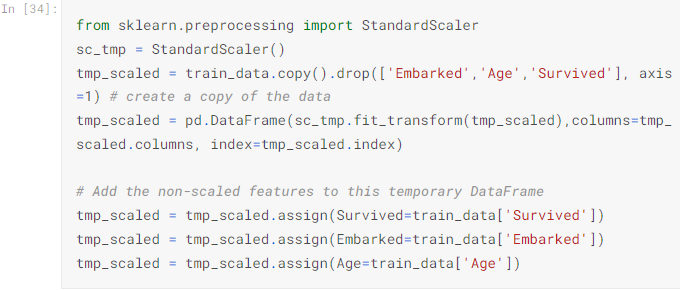
## **4.1. Encoding**

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.



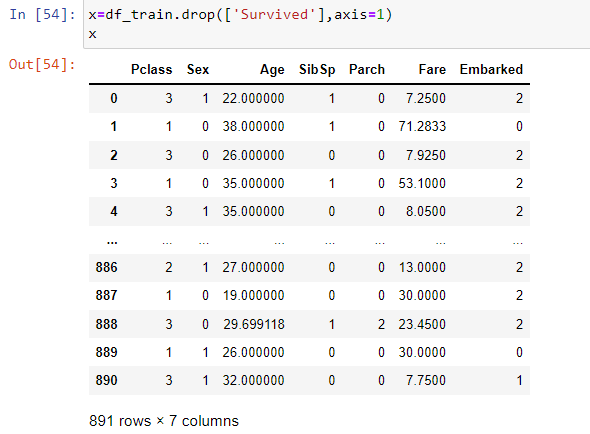
## **4.2 Feature Scaling**

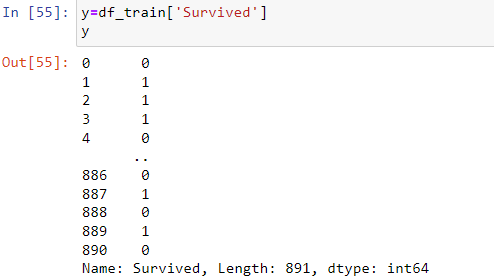
Now we can perform standard scaling on all features except the 'Embarked' and 'Age' feature, because both need to be imputed. This scaling will be performed on a temporary copy of the training data because with the sole purpose of being able to more accurately find nearest neighbors for data imputation. 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:



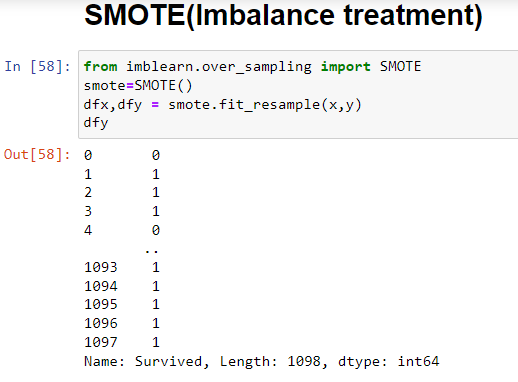
## **4.3 Splitting data into training and testing sets**

Prior to implementation or applying any Machine Learning algorithms, we must decouple training and testing dataframe from our master dataset.

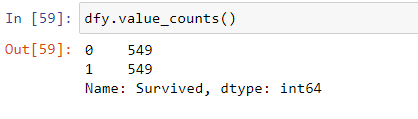




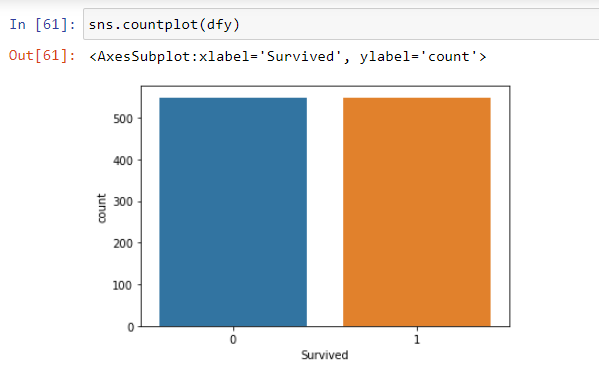
Here we can see our target is not balance we need to balance it.So,using SMOTE we will balance our target



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



Using countplot also we can see that our data is balanced

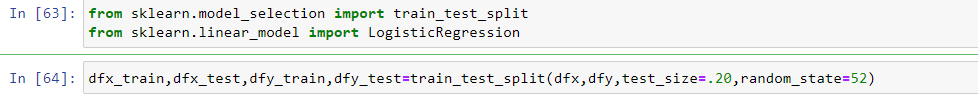
**5.Building Machine Learning Models:**

A machine Learning Model can be a thought of a mathematical equation used to predict a given one or more other values

**5.1 Splitting the Data for Training and Testing**

In ML the separated 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.

It is possible through a inbuilt library of sklearn’s train\_test\_model , it separated the data into test\_size suggested by the DataScientist



5.2 **Training the Models**

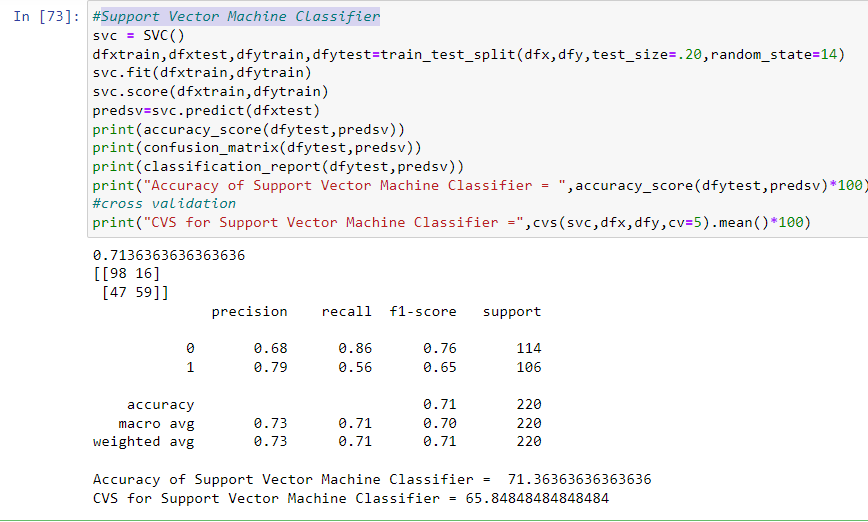
To find the best model it is necessary to train 3-4 models,I have selected 4 models which are:

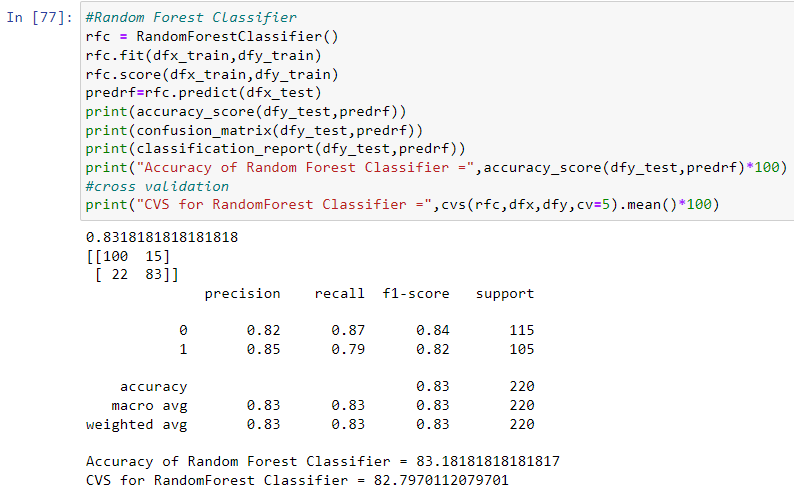
1. Support Vector Machine Classifier

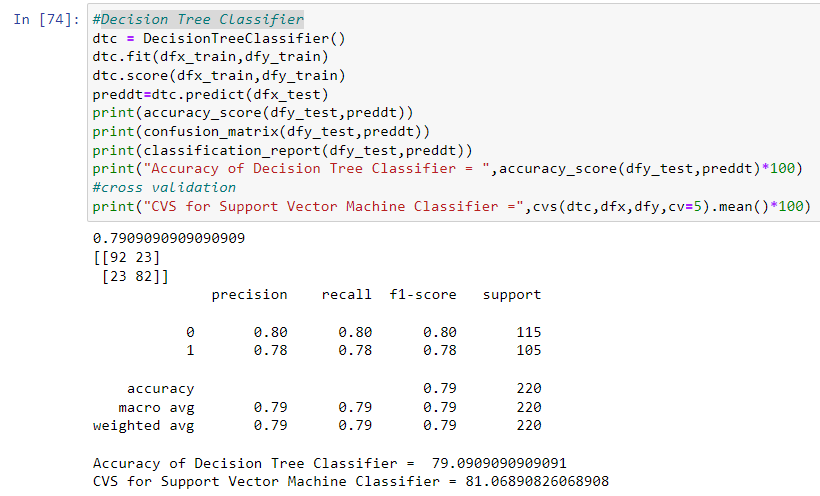
2. Decision Tree Classifier

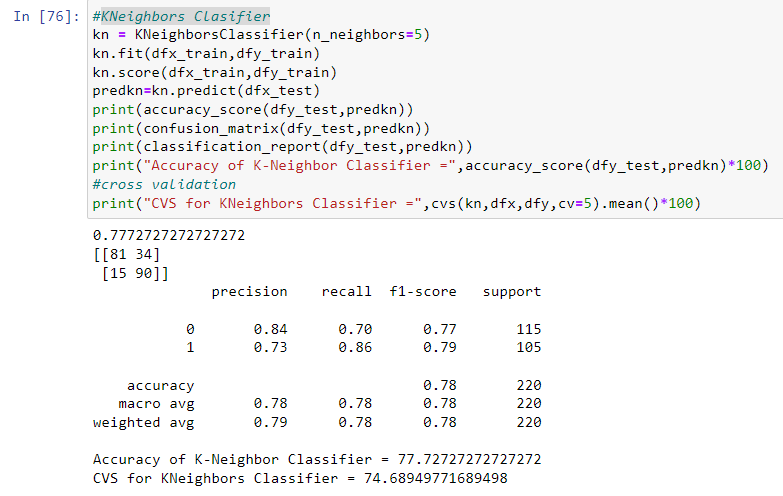
3. KNeighbors Clasifier

4. Random Forest Classifier

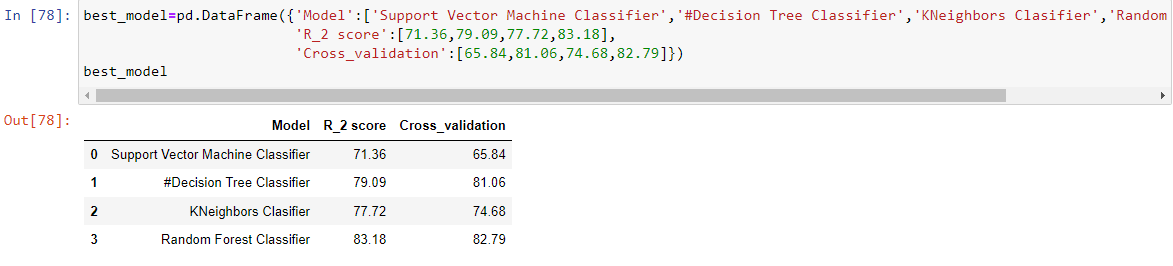








Here I have found Accuracy and Cross Validation of all the four models which I have choosed in this all four We will use Random Forest Classifier because it has the least difference with both accuracy and CV score.



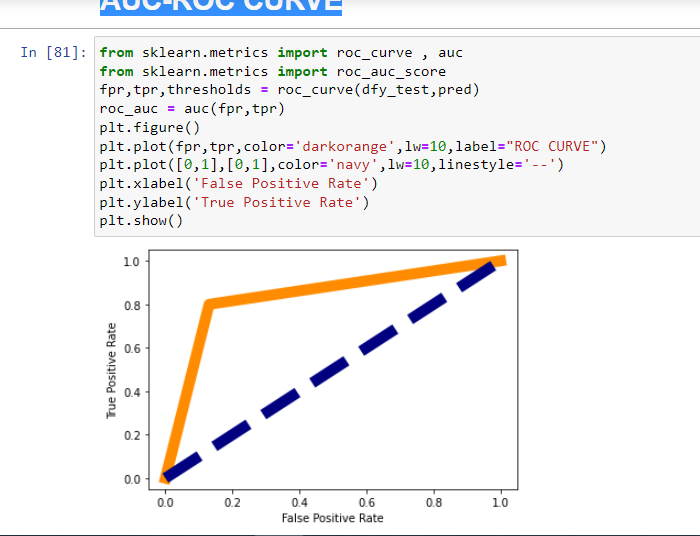
5.3 **Hyper Parameter Tuning Using GridSearchCV**

After selection of model it is necessary to Give the best parameters to the model and it is possible with GridSearchCV method which chooses the best parameters and makes the model fit for exporting



So,finally with th help of hyper tunning test,we can say that our Model is 83% accourate

# Now here we will see the AUC-ROC CURVE



Hence,we have solved missing values and class imbalnce.So, hence we can go with our model.

5.4 **Exporting the Model**

After GridSearchCV the model is exported into .obj file or .pkl file it can be done through Joblib Library or Pickle Library



***6. Conclusion***

So in all and all this was a very exciting projects which gave me a very comprehensive idea about the life cycle of any ML or data science project.It has given me a wonderful insight on how basic math and statistics concepts are used here to solve real world problems.The more and more I read about Machine learning,the more it fascinates me and I am looking forward to apply other algorithms of ML like Decision Tree,Random forest,Support Vector machines.