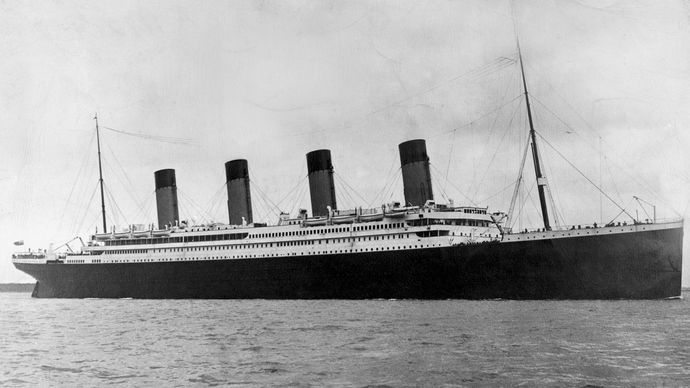
Titanic, in full Royal Mail Ship (RMS) Titanic, British luxury [passenger liner](https://www.britannica.com/technology/ocean-liner) that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history.

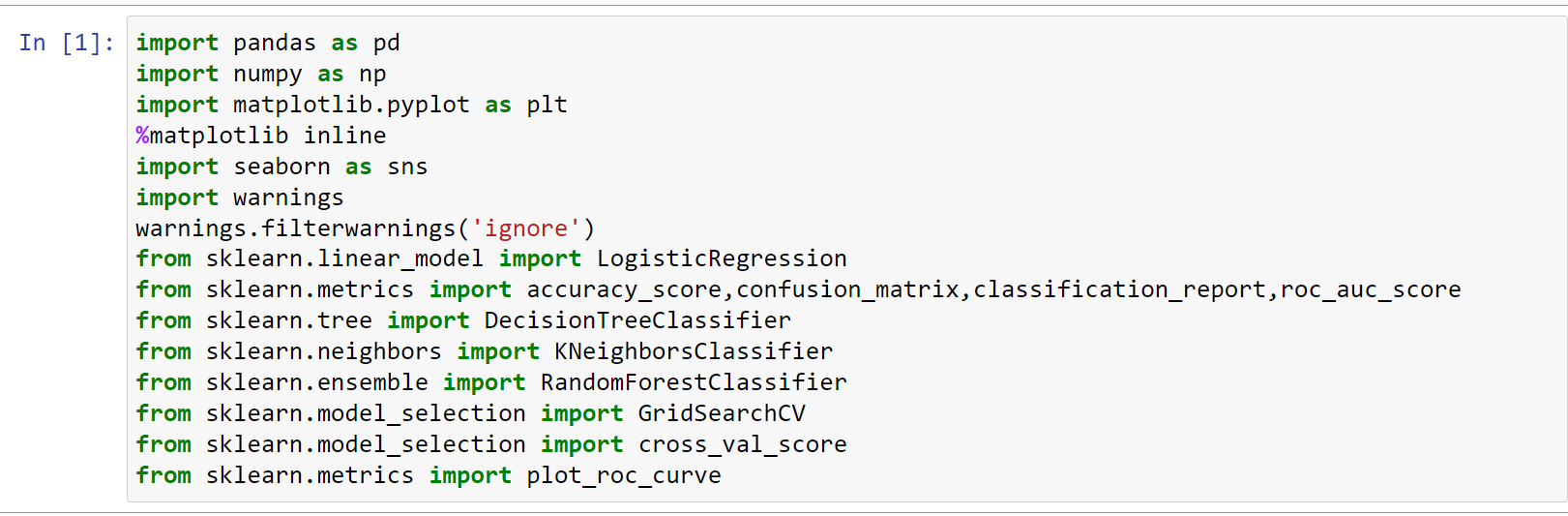


1. **Problem Definition**

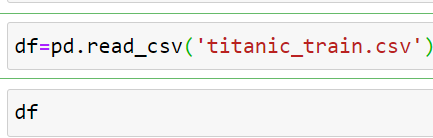
The Titanic Problem is based on the sinking of the ‘Unsinkable’ ship Titanic in early 1912. It gives us information about multiple people like their ages, sexes, sibling counts, embarkment points, and whether or not they survived the disaster. Based on these features, we have to predict if an arbitrary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

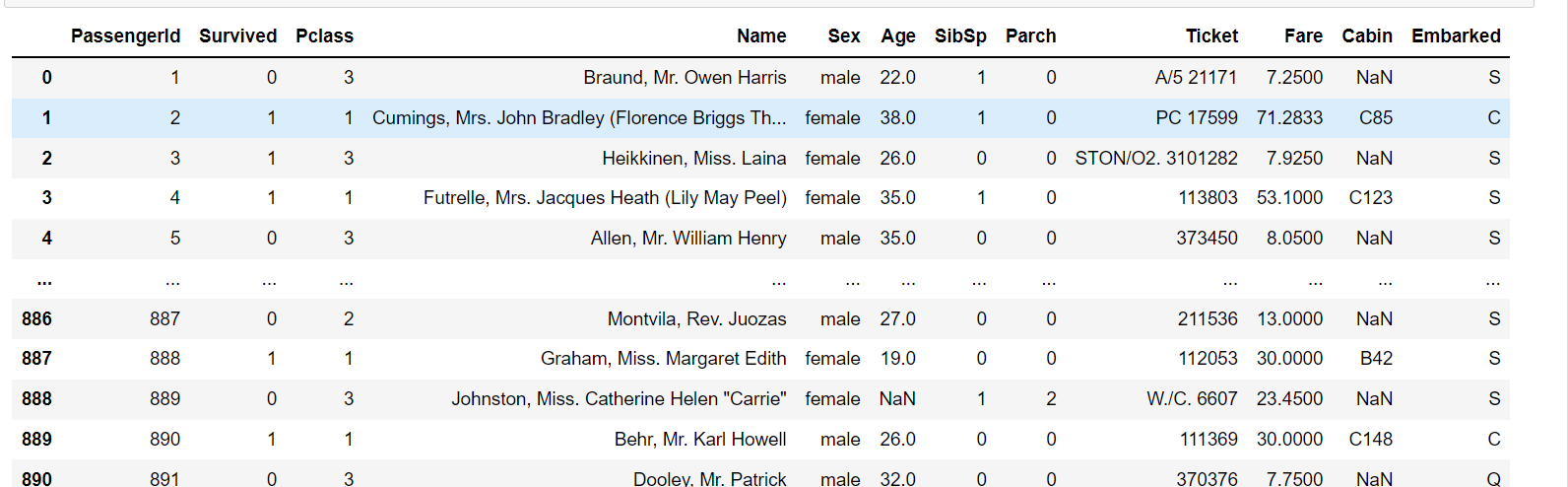
The goal of this project is to build a model that can predicts which passengers survived the Titanic shipwreck.

**Importing the Libraries**



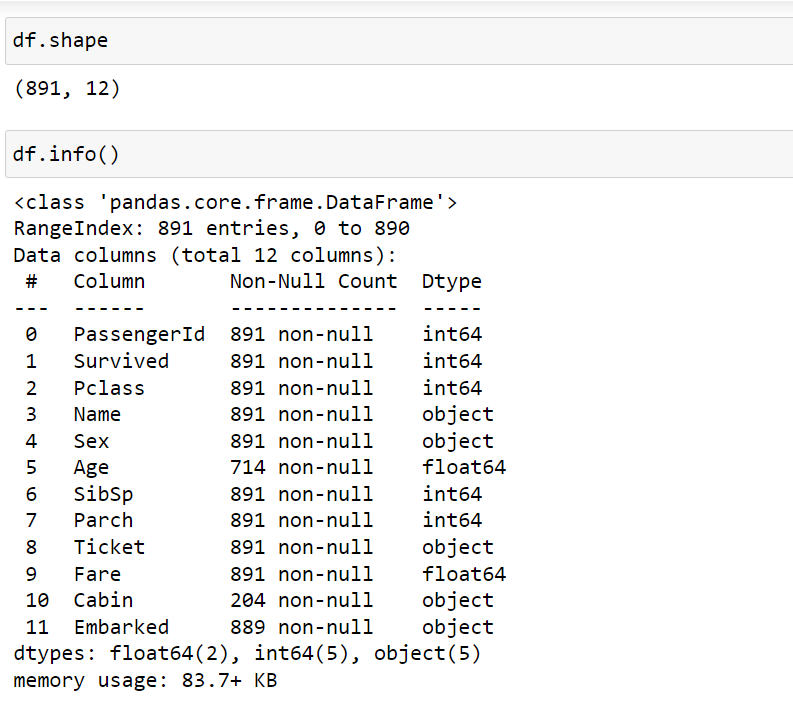
**Getting the Data and Dataset**





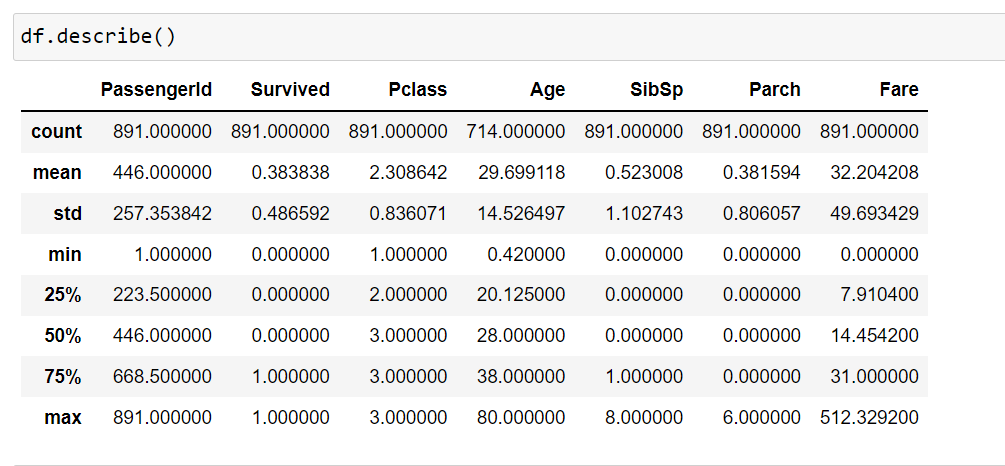
1. **Data Analysis**

In this project, we have a dataset which has the details of the passengers. From these details we have to predict if a passenger will survive or not on the titanic.



The given dataset contains 891 rows and 12 columns. The column names like Passengerid, Pclass, Name, Sex, Age, Parch, SibSp, Ticket, Embarked etc. Survived is our target variable. Here 2 columns are float64, 5 columns are int64, and remaining 5 are object type dataset. Below I have listed features with short description.

For continuous data

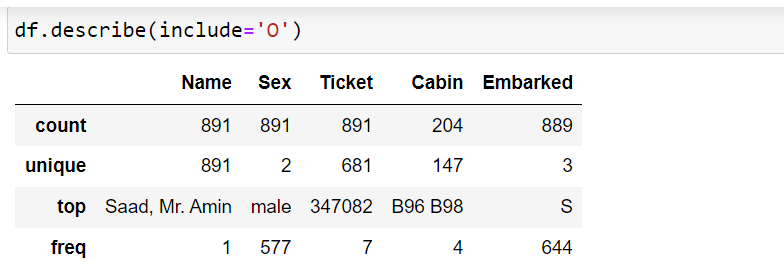


As we can see the dataset contains 891 entries which means 891 passengers were present but if you compare the total number of actual passengers, it was given that around 2,224 passengers that means our dataset is not complete it is just the sample from the actual data

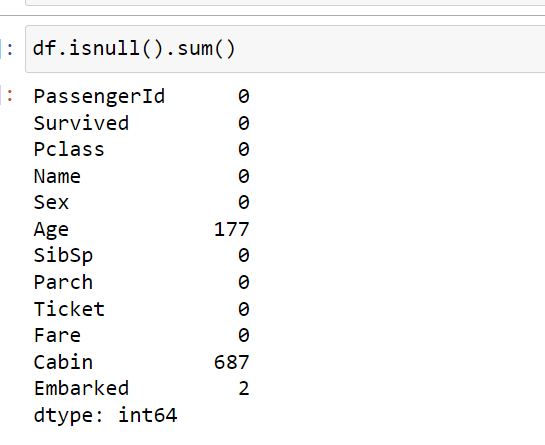
The mean age is 29.699 and the oldest passenger in this data set was 80 years old, while the youngest was only .42 years old (about 5 months).

Here max Fare of a passenger paid for a ticket in this data set was 512.3292 British pounds, and the minimum Fare was 0 British pounds.

For categorical data



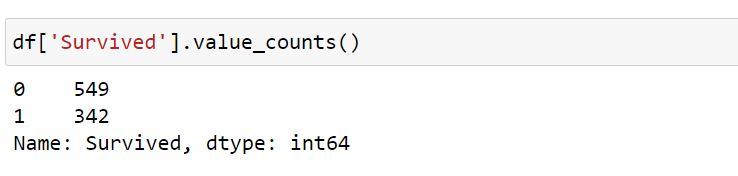
We need to convert a lot of features into numeric, so that the machine learning algorithms can process them. Here, we can see that the features with widely different ranges, that we will need to convert into roughly the same scale. We can also see some features, that contain missing values (NaN = not a number), that we will deal later



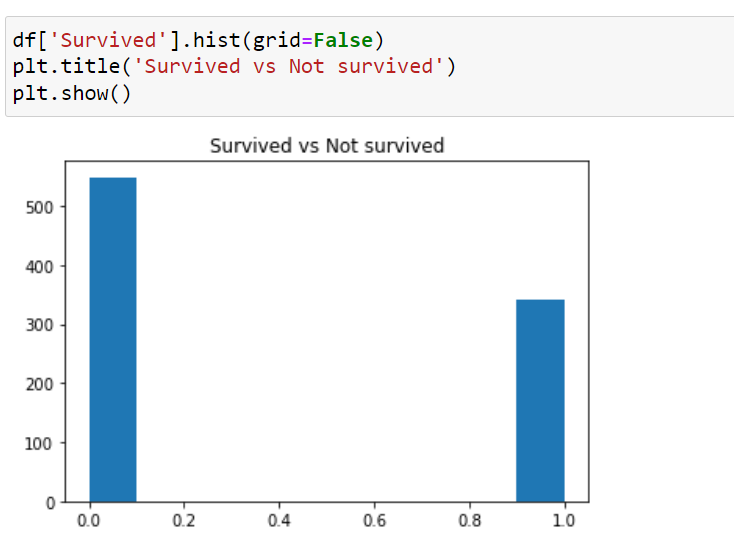
We can see Age, Cabin and Embarked having null values. (Age (714) and Cabin (204), Embarked (889) have non-null count we have to take care of those.

Target variable

Get a count of the number of survivors on board the Titanic in this data set. Notice that, in this data set, there were more passengers that didn’t survive (549) than did (343)



Visualize the number of survivors on board the Titanic in this data set.



1. **Exploratory Data Analysis**

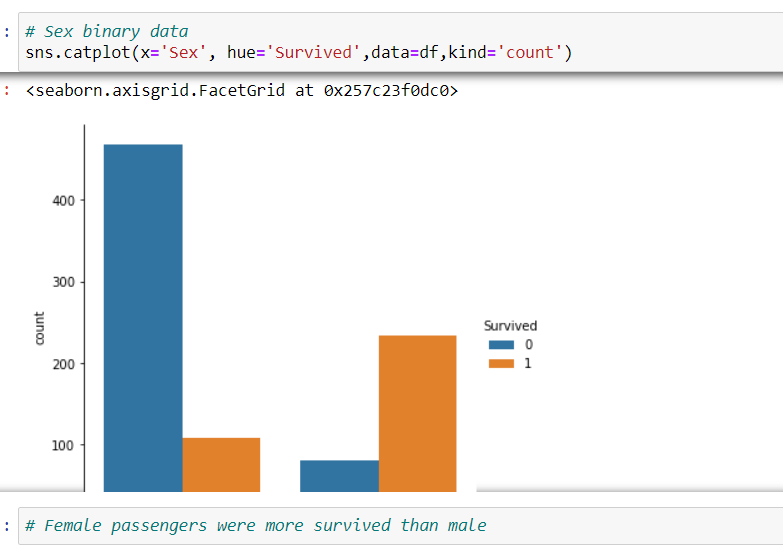
**(EDA) concluding Remarks**

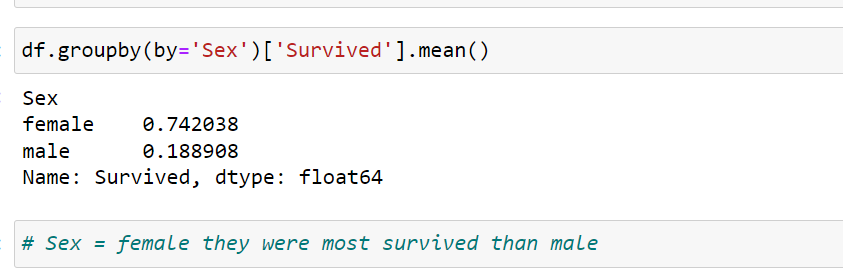
**Dependent variable:** Exploratory data analysis was conducted with the dependent variable, Survived. There were 342 passengers were survived and 549 were not. 38% of the data were survived while 61% were not-survived data.

**Visualizing variables:**

Sex: Female passengers were most survived than men about 74% female

passengers were survived





Age: Children were more survived

Male passenger in between age group (15 to 35) were less survived

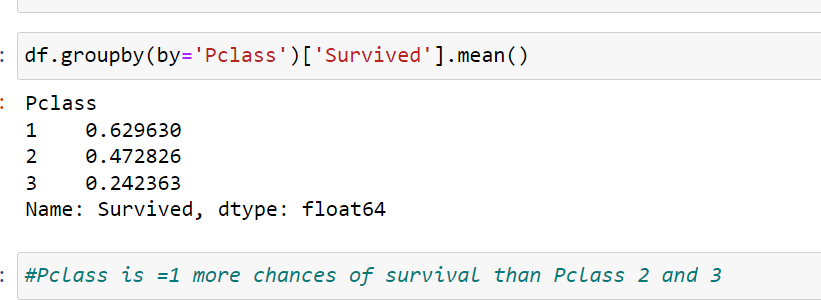
Female passenger in between age group (15 to 40) were more survived

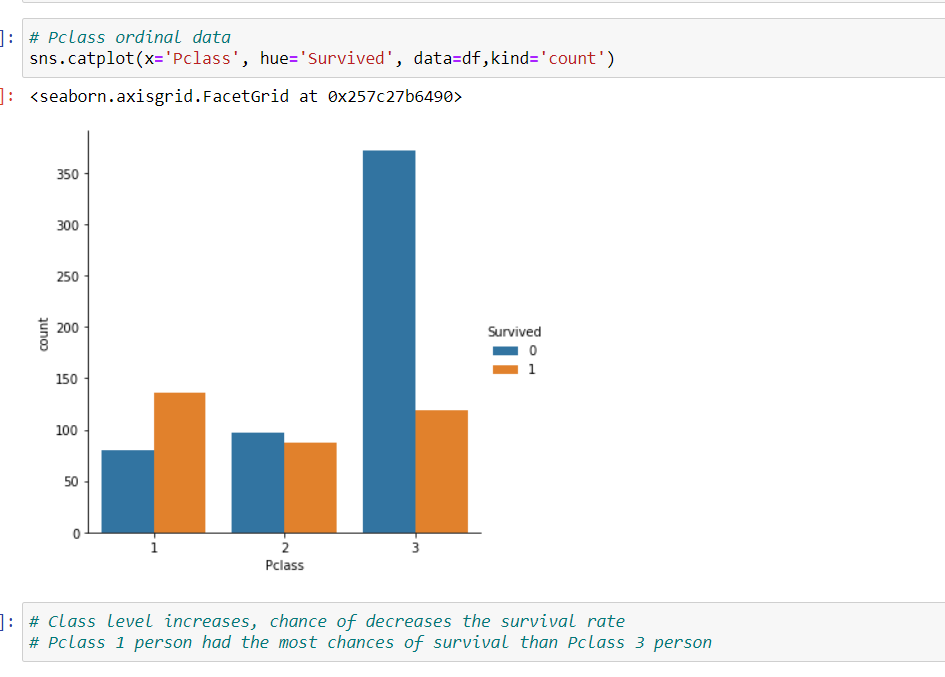
Elders were not survived



Pclass: It is also clear that the men or women who are present in the

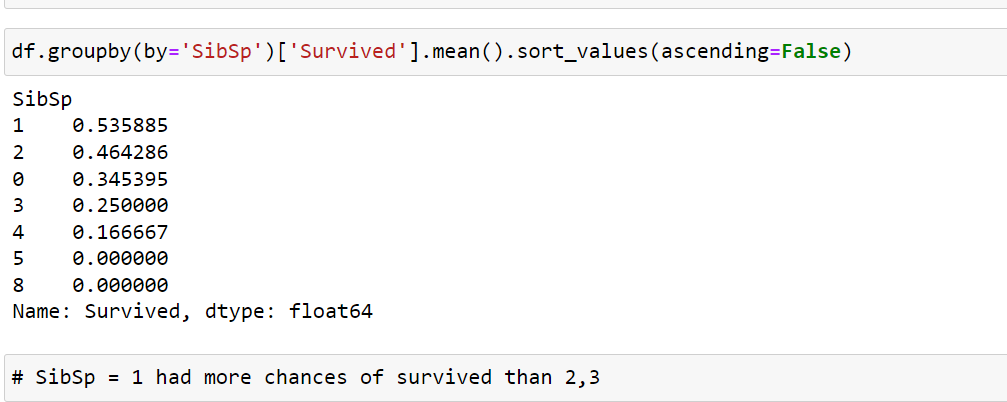
Pclass 1 and 2 have the highest probability of survival.

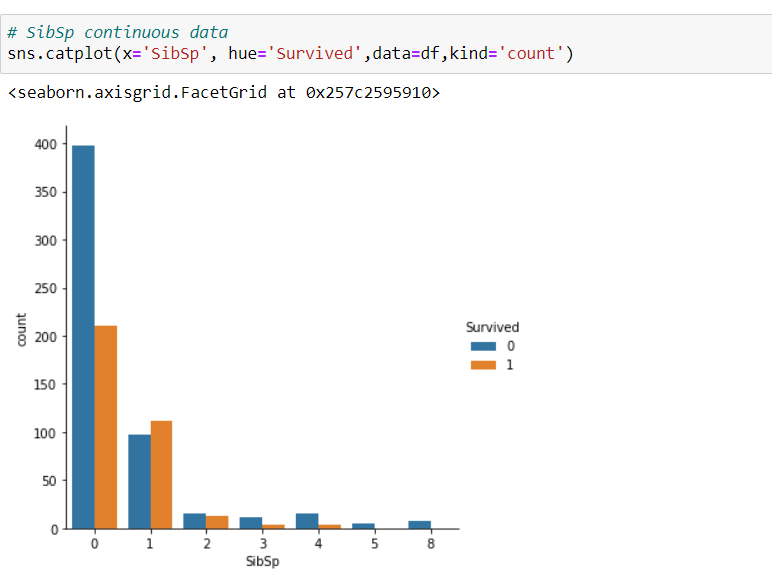




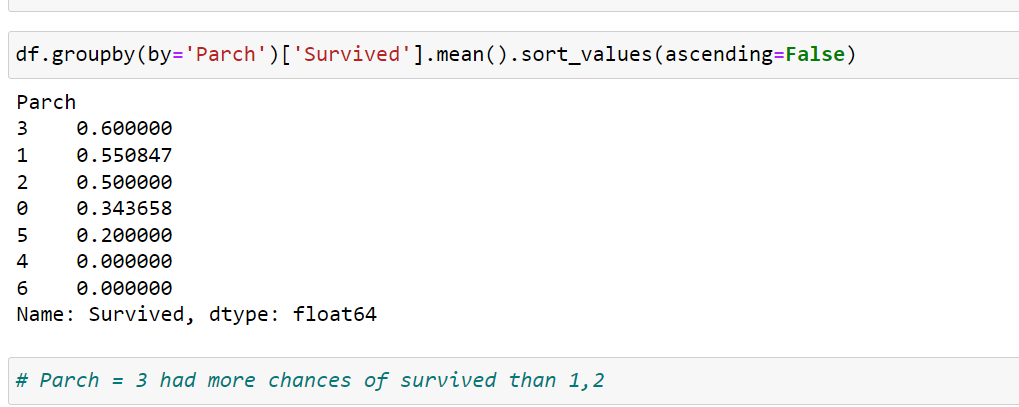
SibSp: Here we can see that the chances of survival decrease as the

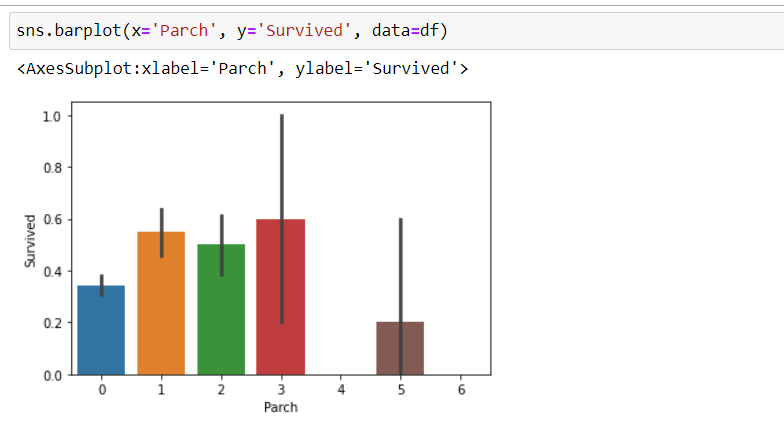
number of Siblings/Spouses increases

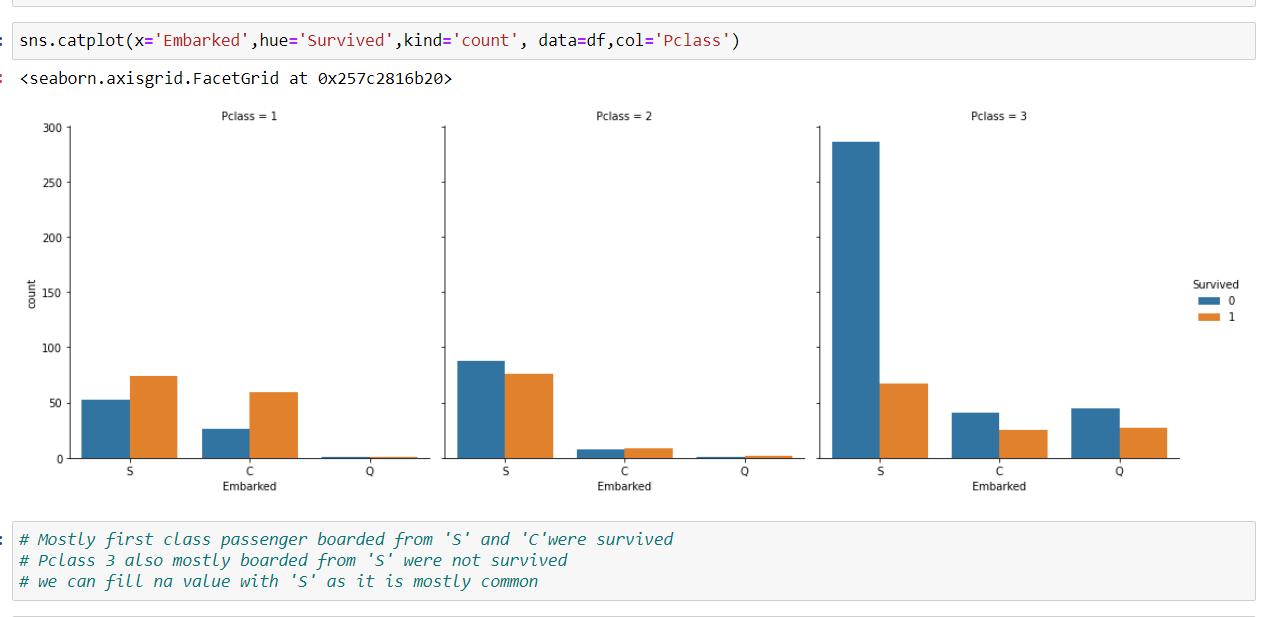




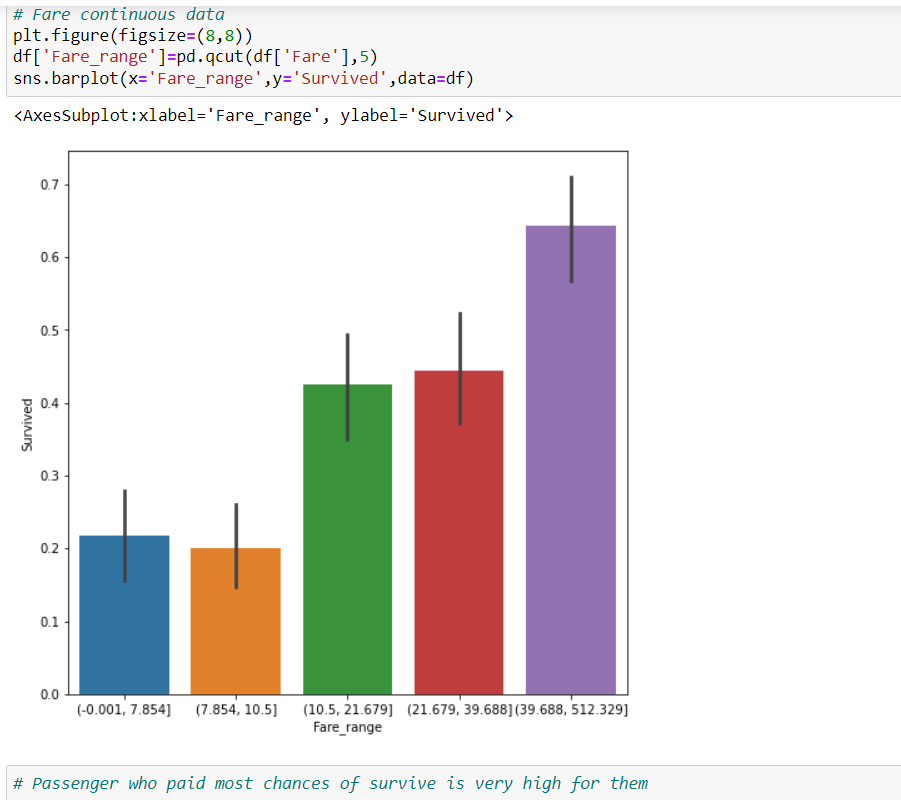
Parch: We can see that the person who has 3 Parents/children with them has the highest chances of survival.



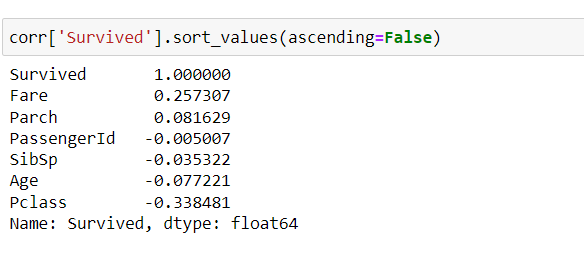




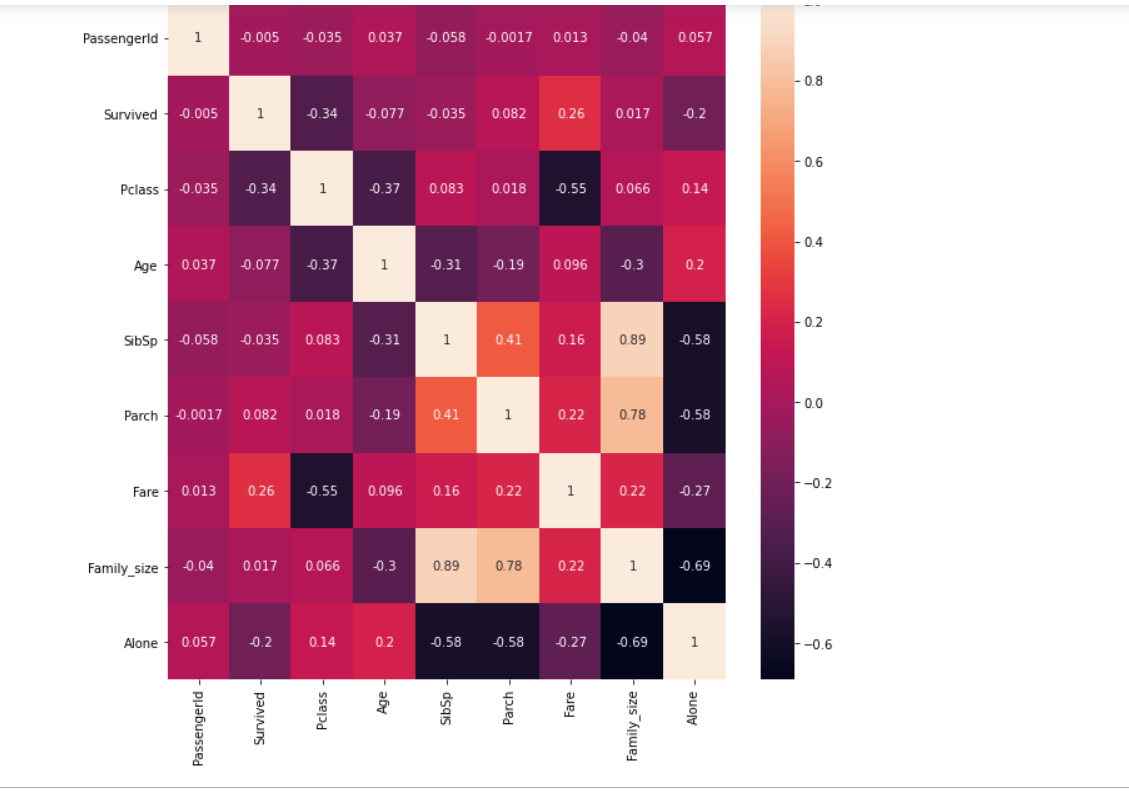
Passenger who paid more survival chances were more



**Correlations among variables:**Heat-map was plotted. By heatmap we can checked correlation between target variable and features



Target variable is negatively correlated with SibSp, Age, Pclass



# Pre-processing pipeline

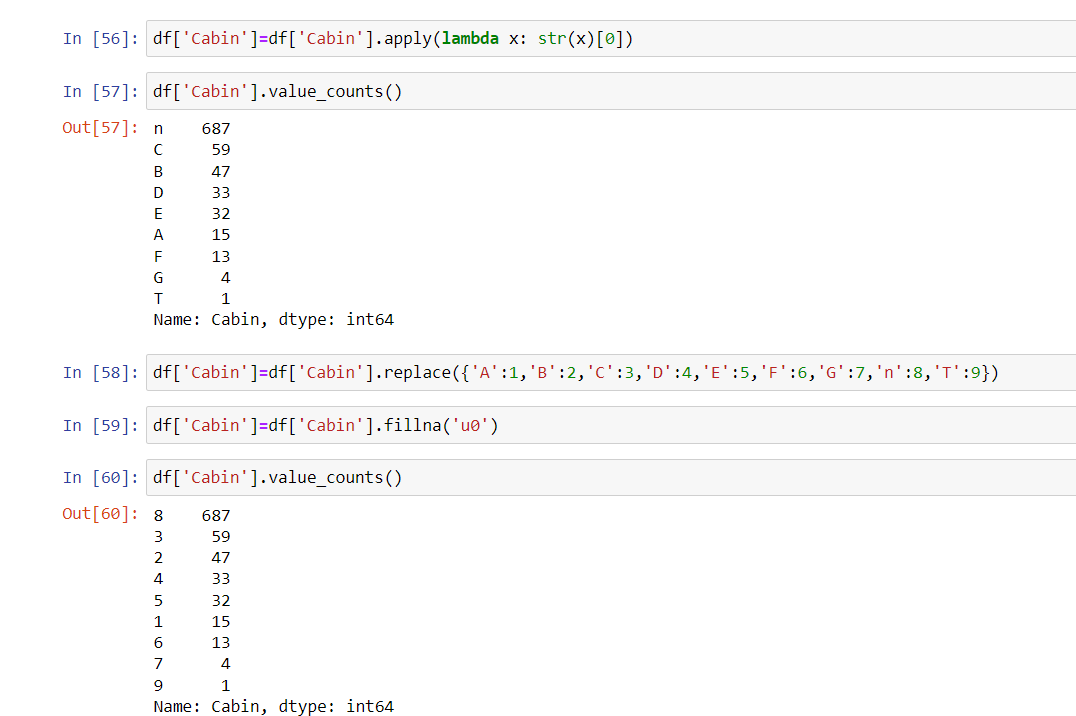
Data pre-processing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, the greater is the reliability of the produced results. Incomplete, noisy, and inconsistent data are the inherent nature of real-world datasets. Data pre-processing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise, and resolving inconsistencies.

* **Incomplete data** can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and malfunctions.
* **Noisy data** can occur for a number of reasons (having incorrect feature values). The instruments used for the data collection might be faulty. Data entry may contain human or instrument errors. Data transmission errors might occur as well.

There are many stages involved in data pre-processing.

* **Data cleaning**attempts to impute missing values, removing outliers.
* **Data integration**integrates data from a multitude of sources into a single data warehouse.
* **Data transformation**such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurement.
* **Data reduction**can reduce the data size by dropping out redundant features. Feature selection and feature extraction techniques can be used.

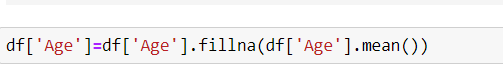
**Treating null values:** Here in the dataset, there are three columns which contain the null values Age, Cabin and Embarked. We will treat all 3 columns

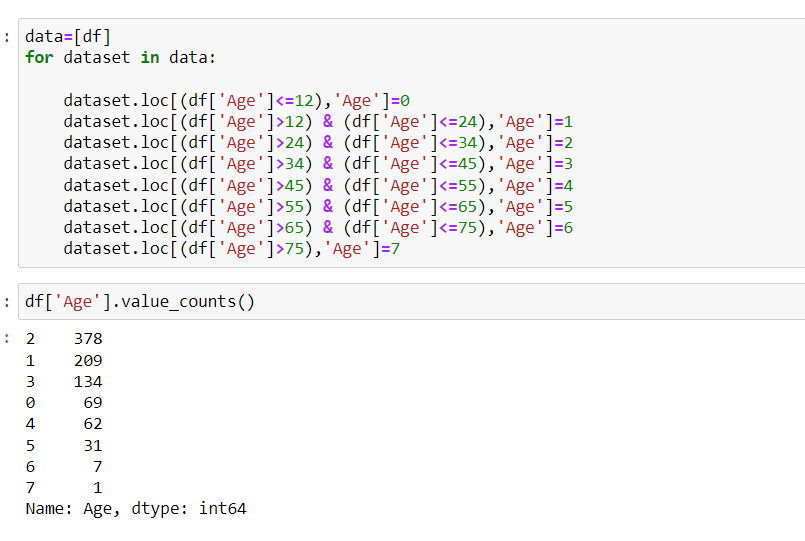
For Cabin

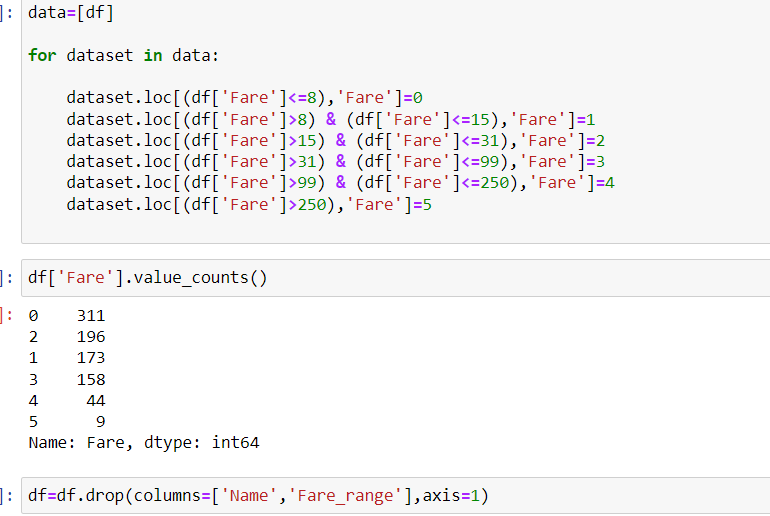
For Embarked



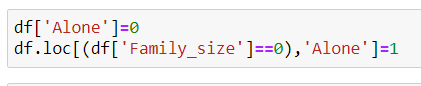
For Age



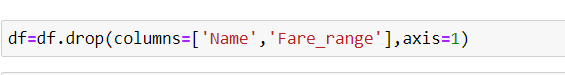
**Dealing with other variables:**



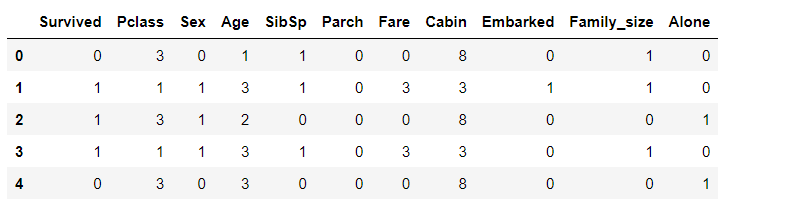




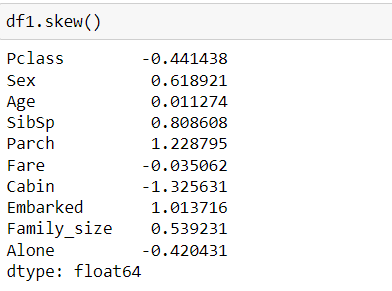
**Dropping Columns:**



**Converting all in numeric**



**Skewness after Power transformation:**

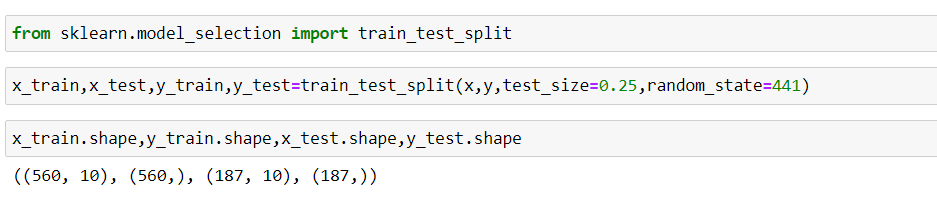


1. **Building models**

For building machine learning models there are several models present inside the Sklearn module.

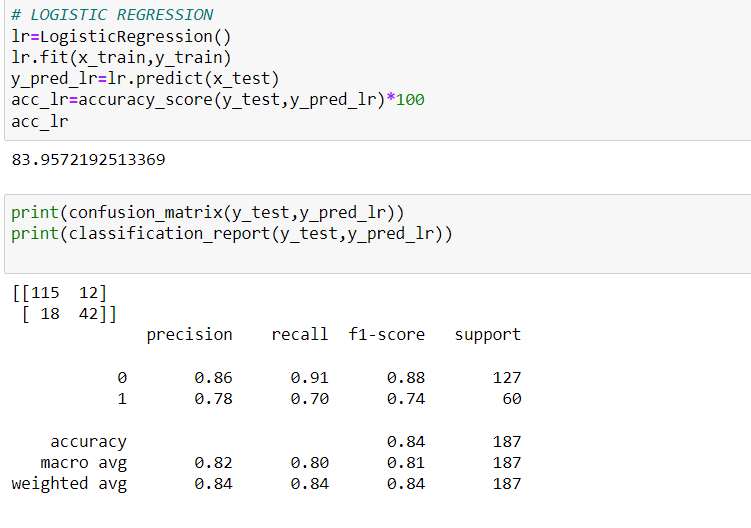
Sklearn provides two types of models i.e. regression and classification. Our dataset’s target variable is to predict whether passenger is survived or not. So for this kind of problem we use classification models.

But before fitting our dataset to its model first we have to separate the predictor variable and the target variable, then we pass this variable to the train\_test\_split method to create a random test and train subset.

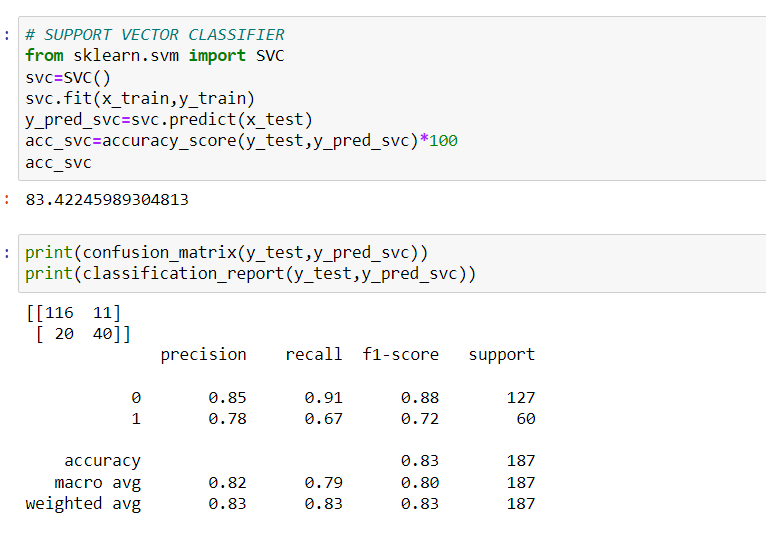


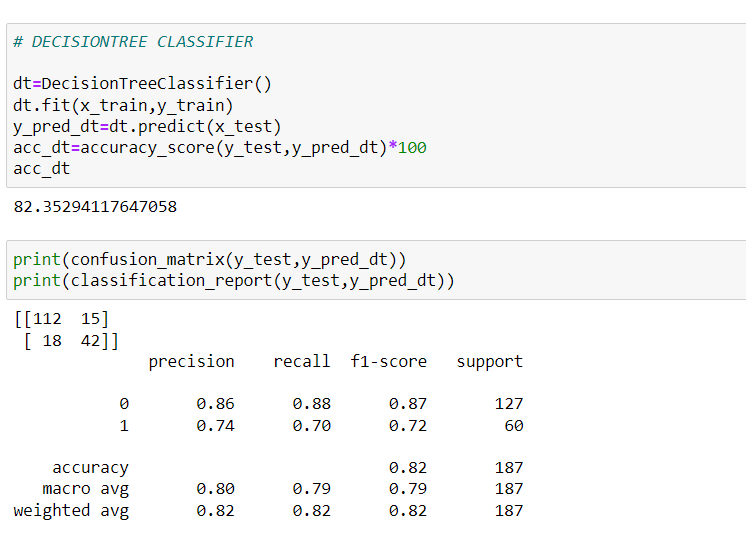
After performing train\_test\_split we have to choose the models to pass the training variable .I have selected 6 models:

**Logistic Regression:**

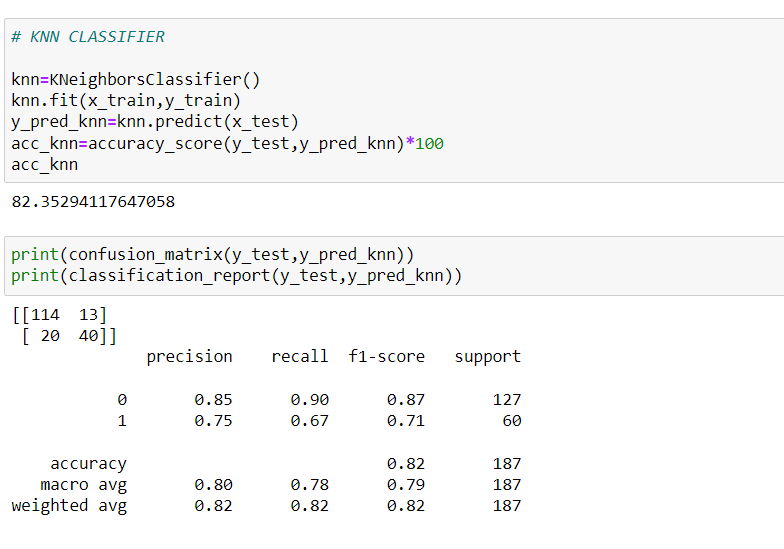


**Support Vector Classifier:**

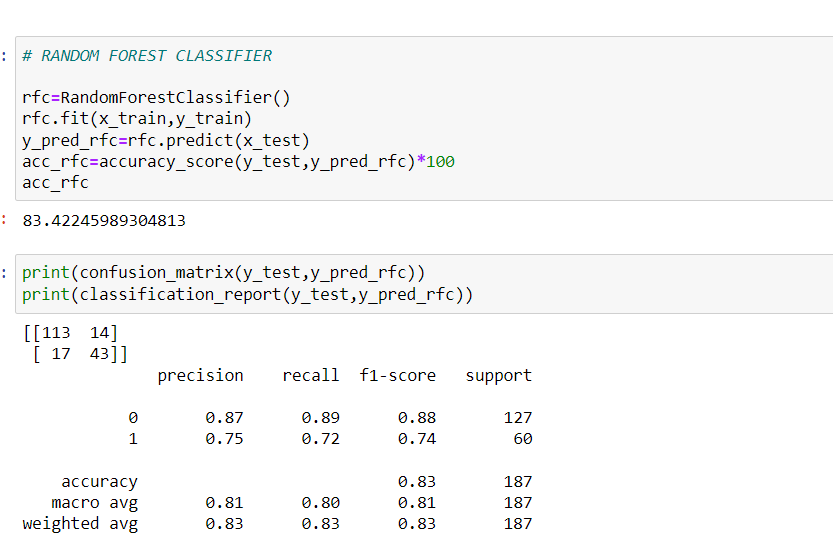
**DecisionTree Classifier:**



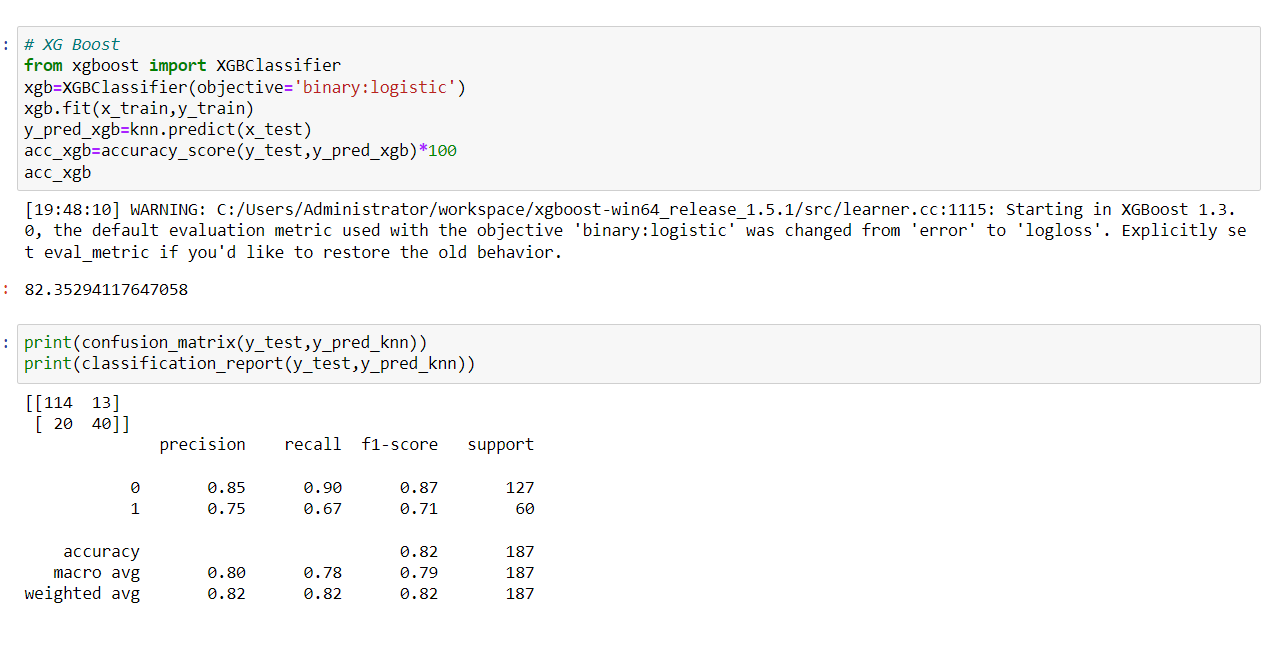
**KNN Classifier:**



**Random Forest Classifier:**



**XGB Classifier:**



1. **Concluding Remarks**

We got our best model i.e. Logistic Regression with the accuracy score of 84%. Here our model predicts 115 true positive cases out of 127 positive cases and 42 true negative cases out of 60 cases.

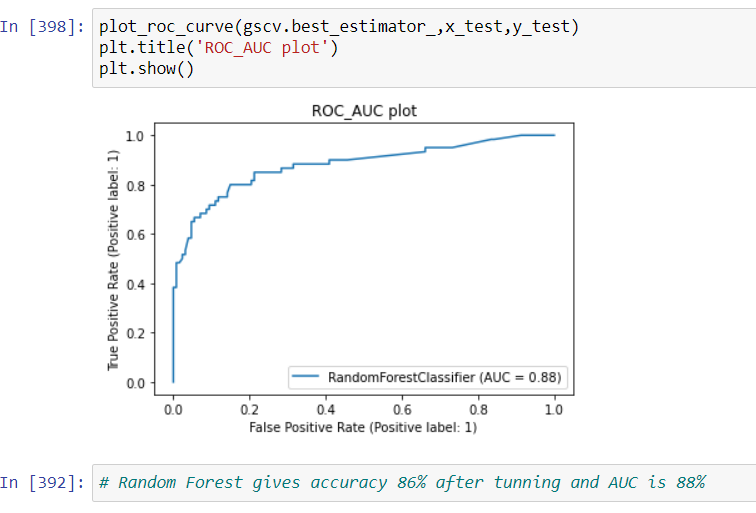
It predicts 12 false positive cases out of 127 positive cases and 18 false negative cases out of 60 cases. It gives the f1 score of 74%.

**Hyper parameter tuning:** Hyper parameter optimisation in machine learning intends to find the hyper parameters of a given machine learning algorithm that deliver the best performance as measured on a validation set. Hyper parameters, in contrast to model parameters, are set by the machine learning engineer before training. The number of trees in a random forest is a hyper parameter while the weights in a neural network are model parameters learned during training. I like to think of hyper parameters as the model settings to be tuned so that the model can optimally solve the machine learning problem.

We will use GridSearchCV for the hyper parameter tuning.



**ROC Curve:**



**Remarks:**

The best and final fitted model was a weighted ***Random Forest*** that yelled a F1 score of *0.74* and a ROC AUC of *0.88*. The model performed excellent. The model’s F1 score and ROC AUC scores were the highest amongst the other models. In conclusion, the model was able to correctly distinguish between passengers who survived and non-survived with high accuracy.

The study is not without limitations. Firstly, this study is restricted by its small sample size. Statistical models are more stable when data sets are larger.

Of course, there is still room for improvement, like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features.