**Kaggle – Titanic Challenge**

*(Vigneshwar K - 270403)*

**AIM:**

To build a Machine learning model to predict which passengers survived the Titanic shipwreck.

**THEORY:**

We are given with two datasets – train.csv and test.csv containing the following information:

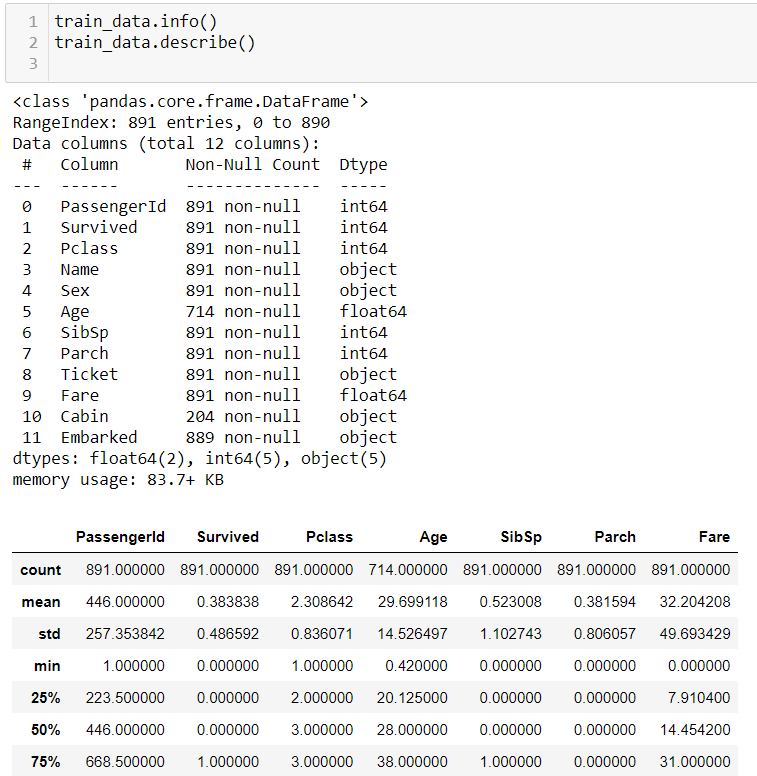
|  |  |  |
| --- | --- | --- |
| Variable | Definition | Key |
| survival | Survival | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| ticket | Ticket number |  |
| fare | Passenger fare |  |
| cabin | Cabin number |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

**IMPLEMENTATION:**

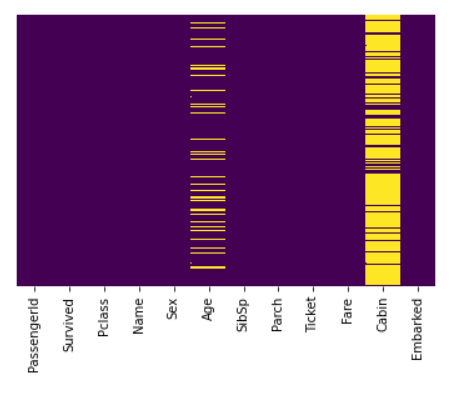
Steps followed:

* Import required libraries and read the csv files.
* Analyse the data in the datasets.
* Pre-process the data in the datasets.
  + Cleaning the data
  + Removing unwanted data
  + Encoding the non-categorical data
* Fit the data into suitable ML model.
* Predict the results of test dataset.

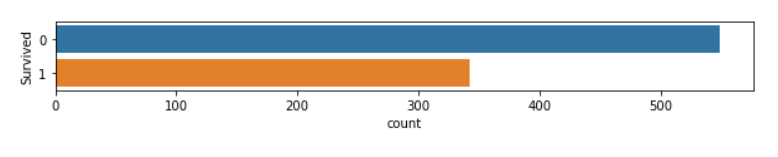
**Analyse Data:**



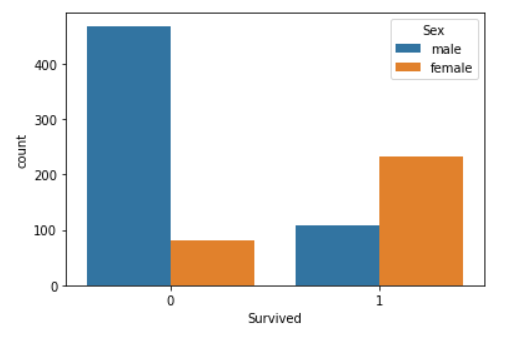
Info() function is used to get a general information of the dataset which includes names and data types of columns. *describe()* function is used to get quantitative data from numerical data present in the dataset. From this we can infer that there are significant number of null values present in Age and Cabin columns. The presence of null values in Age and cabin columns is shown in the heatmap given below. Yellow lines indicate the null values.



**Survivors vs Non-survivors:**

****

**Gender distribution of Survivors:**

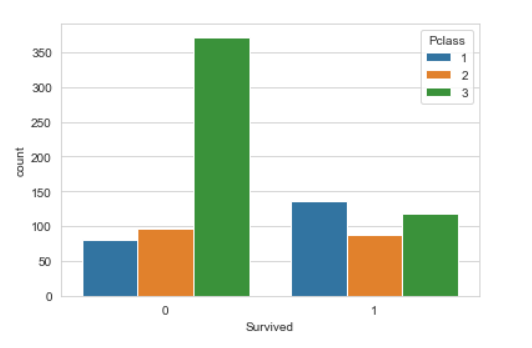
****

% of women who survived: 74.20382165605095

% of men who survived: 18.890814558058924

**With this we can see that only about 18% of men survived and about 75% of women survived**

**Class distribution:**

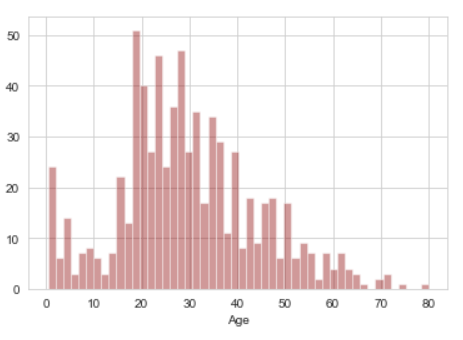
****

% of upper class survived: 62.96296296296296

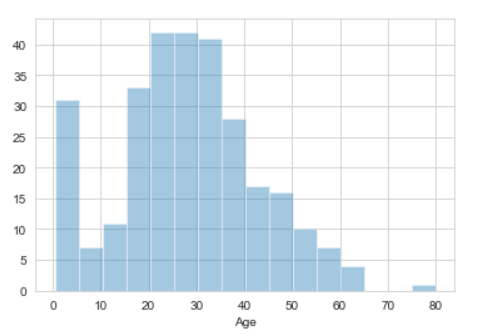
% of middle class survived: 47.28260869565217

% of upper lower survived: 24.236252545824847

**Age distribution of passengers:**

****

**Age distribution of survivors:**

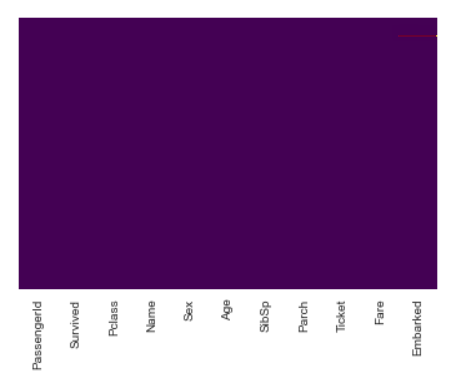
****

From the analysis, it is evident that sex, class and age plays a major role in deciding whether a person survived or not. So, we know our independent variables, next step is to pre-process the data, to make it suitable to fit into the model.

**DATA PRE-PROCESSING:**

1)**Cleaning the data:**

Cabin data is random and it seems to be difficult to fill, so we won’t be considering the cabin data. So, we need to remove the cabin data. But age data is important for prediction, so we fill the null values in the age category with the median of ages in each class (Pclass). Since, there was a significant difference in the average ages in each class attribute and we need both Pclass and Age attributes for prediction, it is good to fill the null values with median of age w.r.t their respective classes.

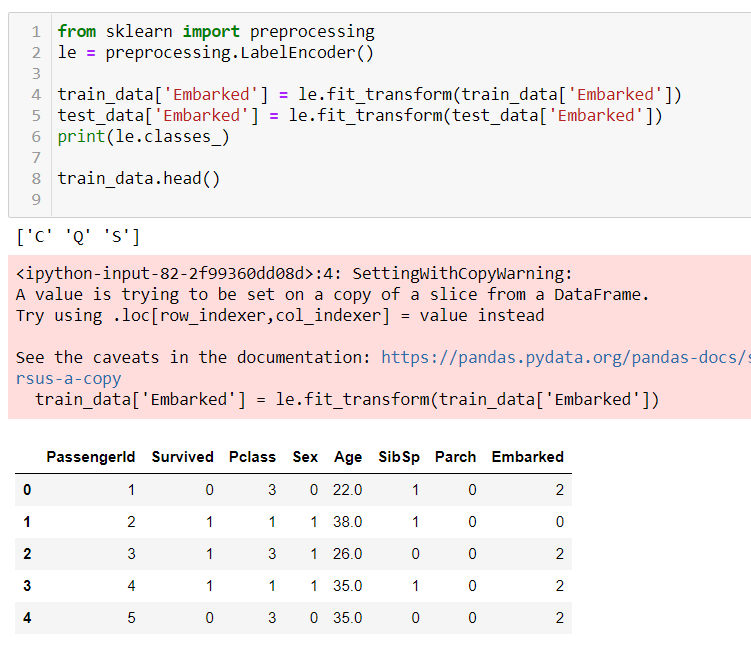


2)**Removing unwanted data:**

Remove Name, Cabin, Ticket, Fare columns from the data. All the above-mentioned data provide little to no use in the prediction. So, it is acceptable to remove the unwanted data.

3)**Encoding the non-categorical data**:

Since, it is a classification problem and we need to fit the data into our model (Logistic Regression), we need to encode the non-numerical data into numerical data. So that we can fit the data into our model. So, Embarked column is encoded using LabelEncoder.



**PREDICTION OUTCOMES:**

Once the data is processed, it is time to fit the data into our model. As it is classification problem, we need to use classification algorithms. Thus, we chose to test with 3 of the popular classification algorithms.

Logistic Regression – 79% accuracy

Decision Tree – 76% accuracy

Radom Forest classifier – 82% accuracy

Random Forest Classifier proved to be best among the three. Further tuning can be done to achieve higher accuracy.