MACHINE LEARNING

FINAL REPORT



**Project Title: Titanic: Machine Learning from a disaster.**

**Course In-Charge: Dr. Nidhi Chahal**

Submitted by

Padegal Sai Giriraj

Katragadda Akash

Chakka Rajesh

**Project Idea:**

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One reason that the wreck prompted such death toll was that there were insufficient rafts for the passengers and crew. In spite of the fact that there was some component of luck involved in surviving the sinking, a few gatherings of individuals will probably make due than others, for example women, children and the upper class.

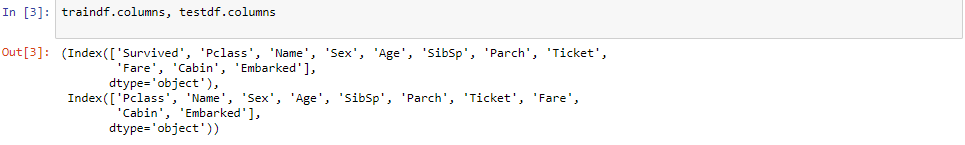
We want to finish the analysis of what sorts of individuals were probably going to survive. In particular, we will apply some machine learning techniques to predict which passengers survived the tragedy and select the classifier which has the highest accuracy.

**Data Set:** <https://www.kaggle.com/c/titanic/data>

**IMPORTING LIBRARIES AND LOADING DATA**

****

**Img-1**



**Img-2**

There are many columns(features) in this dataset let us break them down.

**Survived**: Weather the person Survived or not.

**Pclass**: Passenger class indicates the class of that person aboard the ship.

**SibSp**: Shows the number of Sibling/Spouses they had.

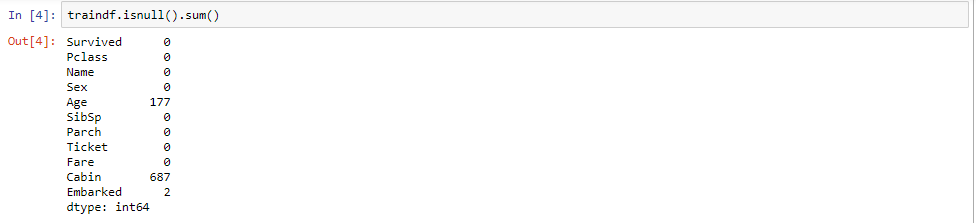
**Parch**: Parch indicates Parents with children

**Ticket**: Ticket name/Number.

**Fare**: How much the Passenger paid.

**Cabin**: Cabin name of that Passenger.

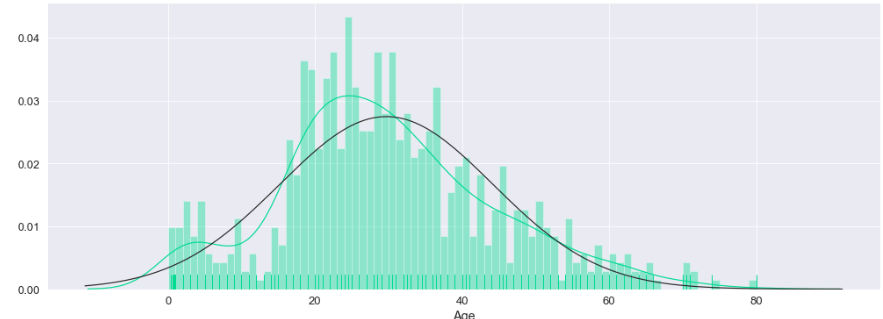
**Embarked**: Point of Embarkation where *C* means Cherbourg, *Q* means Queenstown, *S* means Southampton.



**Img-3**

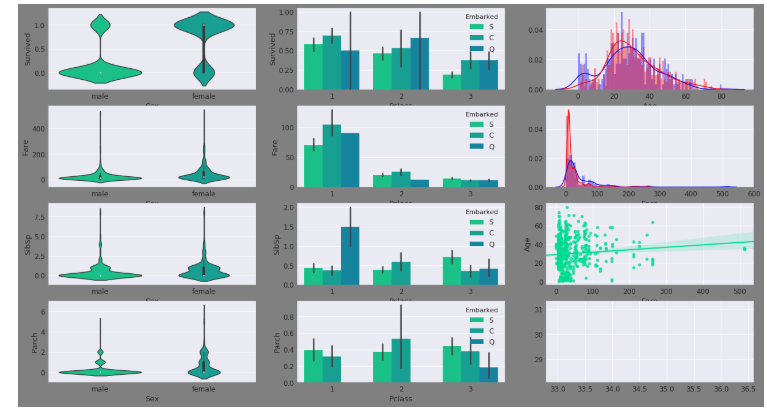
Above Image shows number of missing values in each column

# **Exploratory Analysis**



**Img-4**

Most of the Passengers aboard the Titanic were in the range of 16~ to 40. The age distribution shows bi-modal curve.



**Img-5**

**1.** Starting from the first graph, we can see that very few males survived as compared to female and very few females died in comparison to males.

**2.** First and second class had the most survival rate than the third class whereas passengers who boarded from *'S'* had the least survival rate.

**3.** The *blue* and *red* distribution shows whether the passenger survived or not.

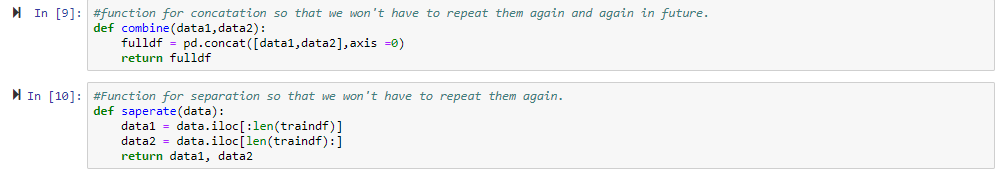
**4. & 5.** There were many passengers from both categories who paid nothing to board the ship particularly from the **third** class

**6.** The distribution of Fare with respect to Survival *blue* indicating Survived while *red* indicating dead.

**7. & 8.** There were more *female* SibSp(siblings and spouses) as compared to *male* and majority of them were from **First** class from *Q* station followed by **third** class.

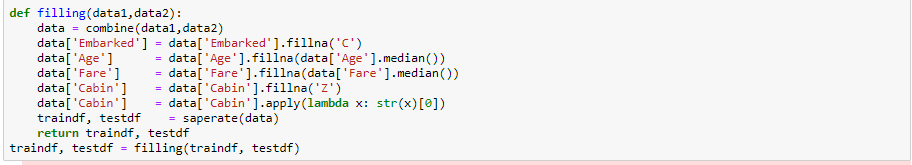
**9.** Distribution of Fare by Age shows that there were many passengers paying nothing being majority while a few paying more than 500!

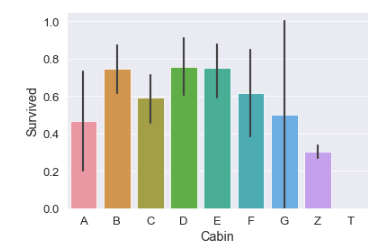
**10. & 11.** By looking at the graph we can see that it is similar with SibSp to some extent with the only difference that *Parch* is flattened and the bar plot says that *Parch* there were no Passengers from *Q* aboard as *1st* and *2nd* class.



**Img-6**

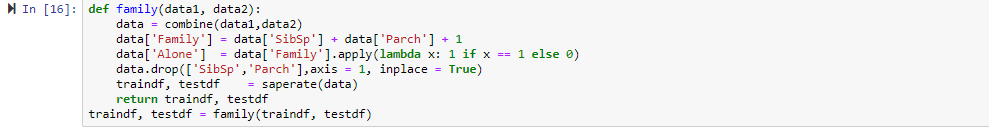
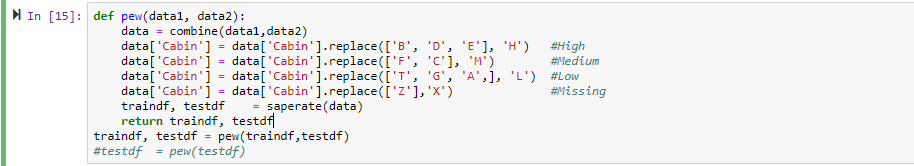
Creating two different functions for combining and separation of data sets





**Img-7**

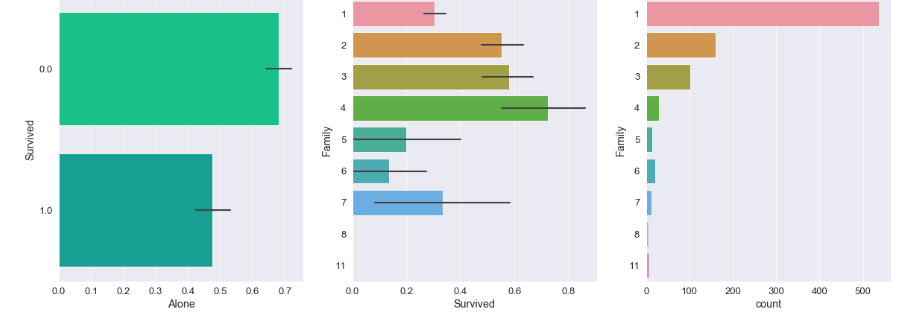
Replacing **Cabin** value with the first character and replacing null value with **Z**. Plot a graph against cabin and survived probability .



**Img-8**

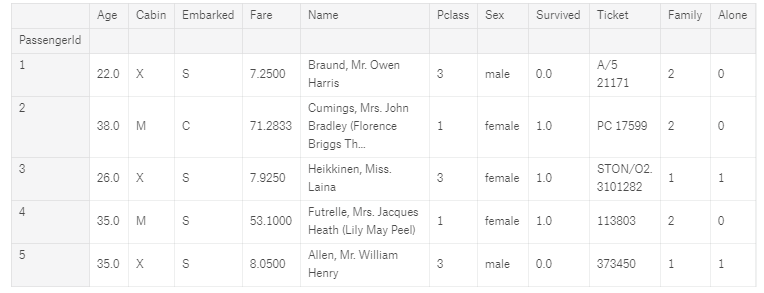
Based on the above graph generated we divided cabins based on their probability into High, Medium, Low, Missing

we combine both SibSp and Parch as they can be represented as family.



**Img-9**

Most of the Passengers aboard were alone. The Passengers who were alone had a lower survival rate which is also true for Passengers who had more than 4 members with them.



**Img-10**

Data after all changes made that are mentioned above

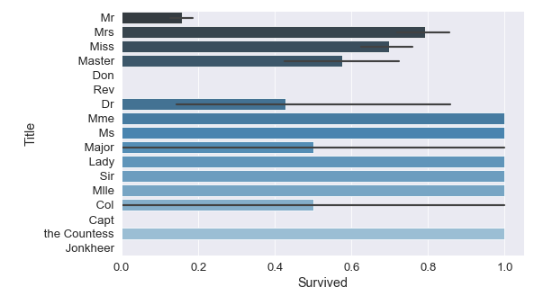


**Img-11**

Assigning some fixed values

If value in **Family** column is greater than 2 changing the value to 3 and if equal to 2 and equal to 1 keeping it same

Discrete values to Embarked, sex and cabin.



Img-12

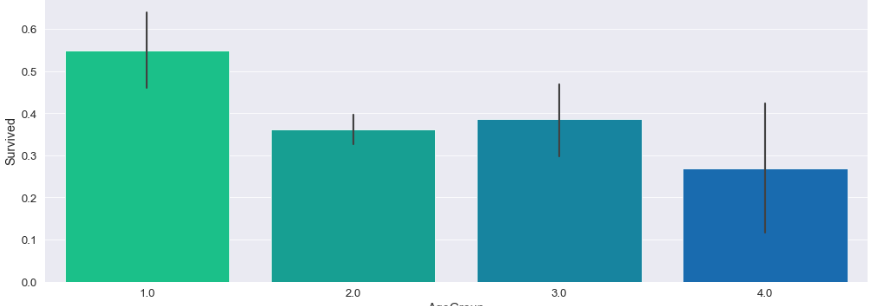
Divided by their names based on their Probability of survival.

'Mme', 'Ms', 'Lady', 'Sir', 'Mlle', 'the Countess' having high probability so they are mentioned as High



**Img-13**

Divided based on their surname

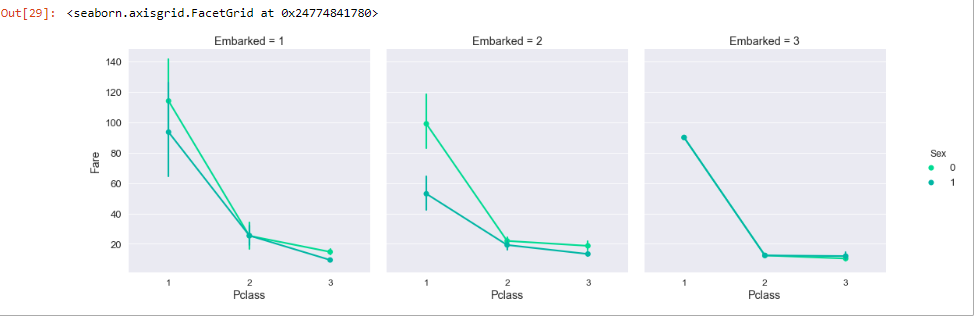


**Img-14**

Making discrete based on their ages like

Age <=16 is 1, Age >16 &Age <=40 is 2

Bar graph shows age vs survival probability



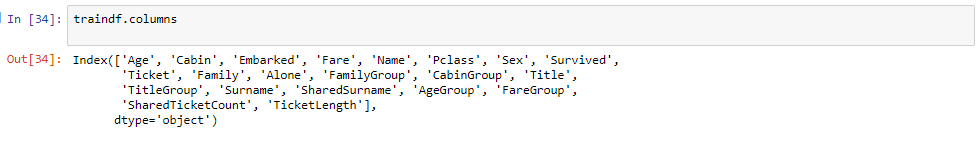
**Img-15**

There is much difference for 1st and 2nd Embarkation for 1st and 3rd Pclass in terms of fare for males and females while the 2nd class fare is similar in all the Embarkations.



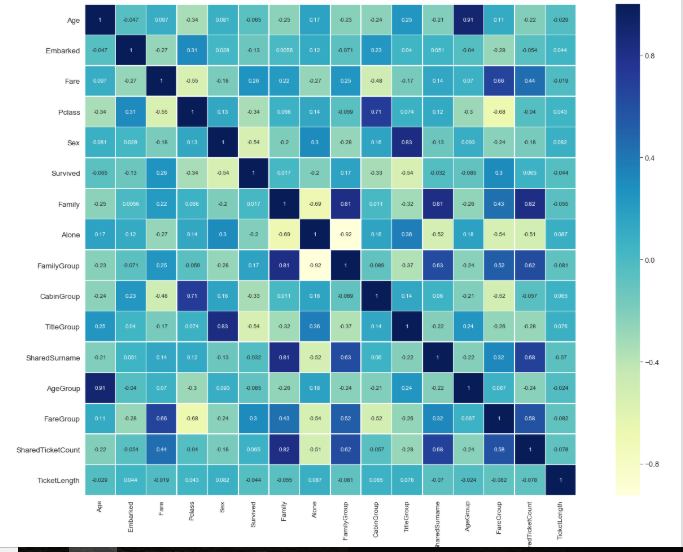
**Img-16**

Grouping Fare and creating a new column called *'FareGroup'* with their means by Pclass



**Img-17**

After all changes made all columns have been displayed



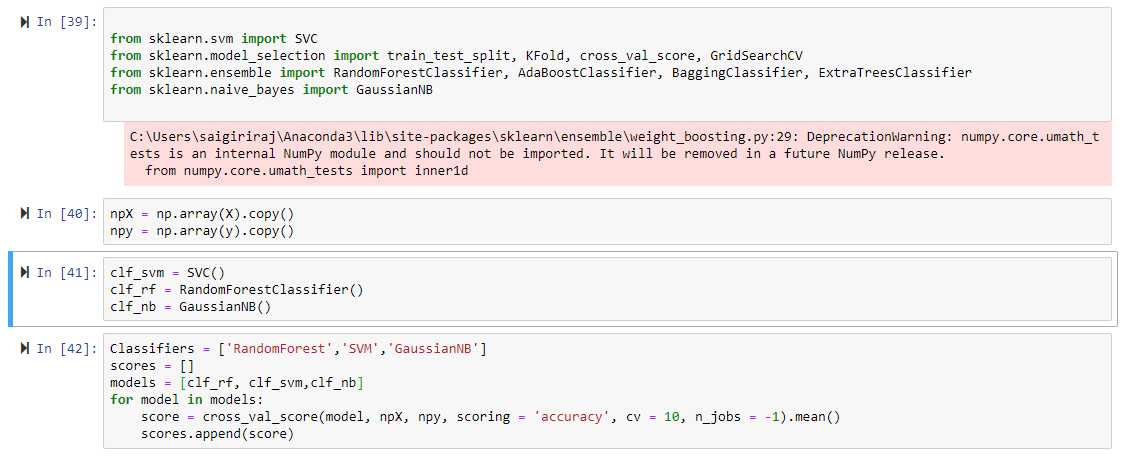
**Img-18**

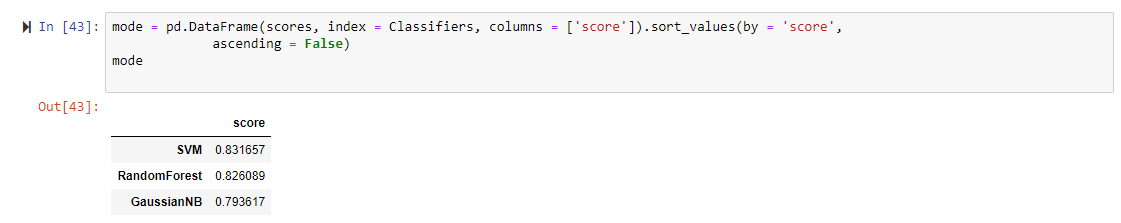
Since categorical features have been created from the features present in the dataset taking only the categorical for training the models.

**MODELS**

We used 3 different classifiers

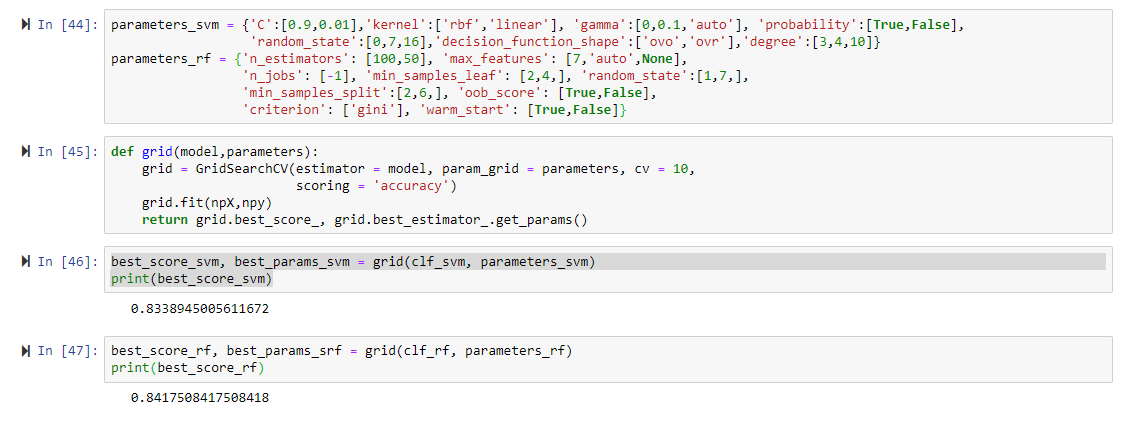
* SVM
* Random Forest
* Naïve Bayes





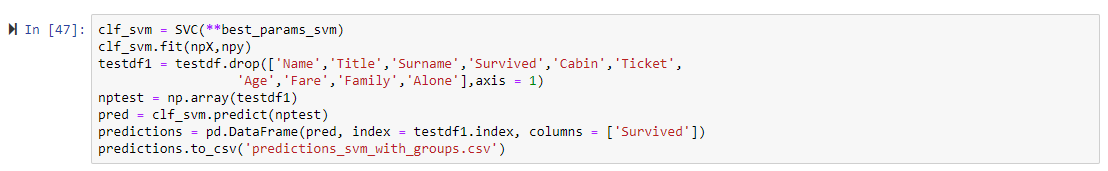
**Img-19**

Accuracy of classifiers before applying parameters



**Img-20**

Accuracy after applying parameters to the classifier

****

Finally storing the result in a CSV file

**Accuracy comparison**

|  |  |  |
| --- | --- | --- |
|  | Before applying Parameters[%] | After applying parameters{%] |
| SVM | 83.16 | 83.33 |
| Random Forest | 82.60 | 84.17 |
| Naïve Bayes | 79.36 |  |

As per obtained accuracy before applying parameters SVM gives more accuracy and after applying parameters Random Forest gives more accuracy**.**