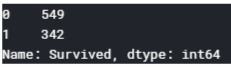
As a machine learning practice project, I took the Kaggle competition of Titanic – machine learning from disaster and worked upon the dataset.

**Data Analysis**

As the name suggest the dataset contains the data of the people travelling through the Titanic and their full details with following columns.



So, we will use machine learning to predict that a person is survived or not as per its data provided. The distribution of people survived or not is shown below.

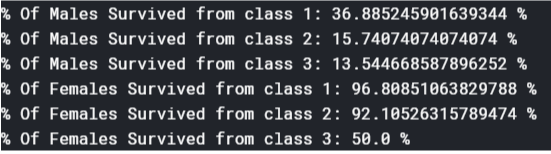


The no. of people survived is less then the no. of people not survived. But as per we have read from the books and shown in the titanic movie we know that the priority was given to females and children for survival. Then we should check the rate of survival for male and female.



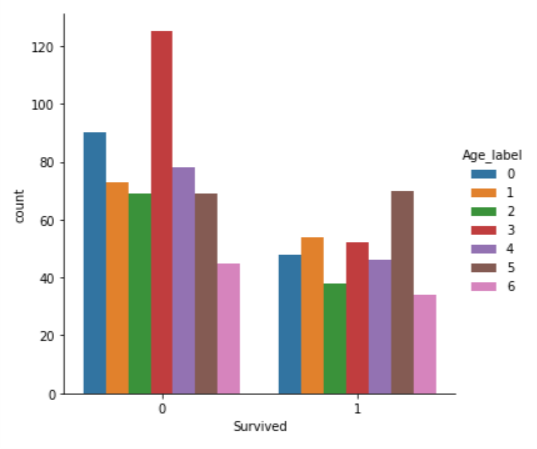
Most off females onboard survived and most of the male onboard could not survive.so the sex will be a good feature in prediction.

Other observation is that the preference was given to the first class over second class and third class.



The survival rate is higher in 1st class then the 2nd and 3rd class. So, the passenger class will also be a good feature.

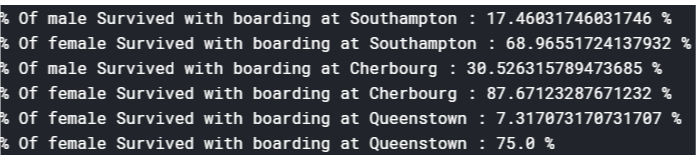
Now to get the insight if the age is also a good feature on not I converted the age column with continuous values into 7 discrete values.



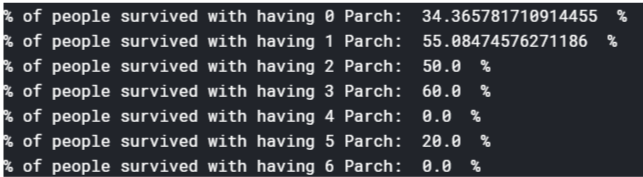
With further research I found that the survival rate was higher in children and the elders. As the survival rate was low in the people in the middle age.

I also converted fare column to discrete but was not much insightful as the higher the fare will mostly suggest the higher class which we have data for.

The boarding of titanic was done at three cities named as Southampton, Cherbourg & Queenstown. The statistics of survival as per the boarding station is given below.

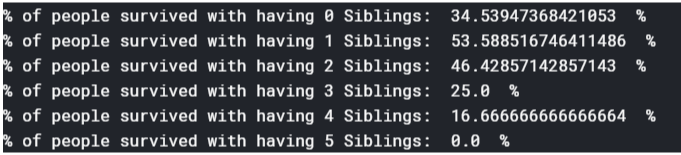


SibSp and Parch columns represent the no. of siblings and no. of parents/children respectively. This can give a great insight because as per the documentation of Titanic the preference was given to the people having children over other.



The above condition can be used only for the Parch up to 3 as we have very less data over it which will be not helpful to us.

The same analysis can be done over SibSp which give us following statistics.



Again, we should not consider it for more then 2 siblings as we have very less data for them.

So, this was insights from data analysis.

**Note: The data analysis done in this project is much more then these, but we can’t include it all in the report. You can see it on my GitHub or Kaggle link.**

**Feature extraction**

The most important part of any machine learning project that gives us chance to showcase our skills and logical thinking. I have converted the Sex column to label where Male=0 & Female =1.

**(Sex\_labels + 1) / Parch**

As seen in data analysis the sex is better parameter for prediction. The female will have high survival chances while the men in same class will have low survival chances. Here +1 is added as we have given 0 to male sex label. Now as the people having children were given preference, we have divided it with Parch. So, Higher this quantity higher chance of survival.

**Parch + SibSp + 1**

As family were given preference it was a good choice to combine attributes that shows the no. of people of family present on ship. To avoid 0 I have given +1 in feature.

**Parch / Pclass**

As having family in higher class will have highest chance of survival then being single in lower class.

**Parch + SibSp + sex\_label**

As being female having parents and siblings with family will increase chance of survival Then being male not having any siblings and parents.

There are also other changes made to features available in Data to convert them in proper format like Age to Age\_label, Fare to Fare\_label, etc.

**After this I used three type of databases for model training.**

* Db1-Regular Data
* Db2-Standard Scaled Data
* Db3-Min Max Scaled Data

**I used all of databases on following models.**

* Logistic Regression
* Support Vector Machines
* Decision Tree Classifier
* Random Forest Classifier
* Neural network

The result is shown on next page.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Name** | **Database** | **Train Accuracy (%)** | **Test Accuracy (%)** | **Model** |
| Logistic Regression | Db3 | 80% | 78% | Model 3 |
| Support Vector Machines | Db3 | 79% | 78% | Model 6 |
| Decision Trees | Db2 | 86% | 83% | Model 8 |
| Random Forest | Db2 | 87% | 84% | Model 11 |
| Neural Network | Db3 | 62% | 60% | Model 13 |

**Conclusions:**

Below are the steps you should consider for being safe our survive from a disaster like Titanic.

* Try to Travel in Higher class rather than lower class if you can as they are given preference and also their rooms are safer then lower class.
* Be female or you can say prefer to travel on ship only if you are female.
* Try to travel with family.

Thank you for giving your valuable time to read my report.