Project Work

Machine Learning to predict Titanic Survivor

Preface

- Goal of the project was to create a Model which can correctly predict the survivor of the Titanic using the machine learning techniques to create a mathematical dependence bw the various properties of the passenger and his Survival.
- The dataset used in the Project is the Kaggle's Titanic Dataset which can be found on kaggle.com/c/titanic .

Parts of Project

The given Project is divided into main three parts:

- Data Visualization
- Data Preprocessing
- Creating a Model

Data Visualization

The process of data pre-processing analysis or data visualization is the process of determining the input parameters for the mathematical Model which makes a high impact on the survival of a passenger.

These chosen parameters then will be used as an input to our Machine Learning model so that the unessential and unwanted parameters can be avoided.

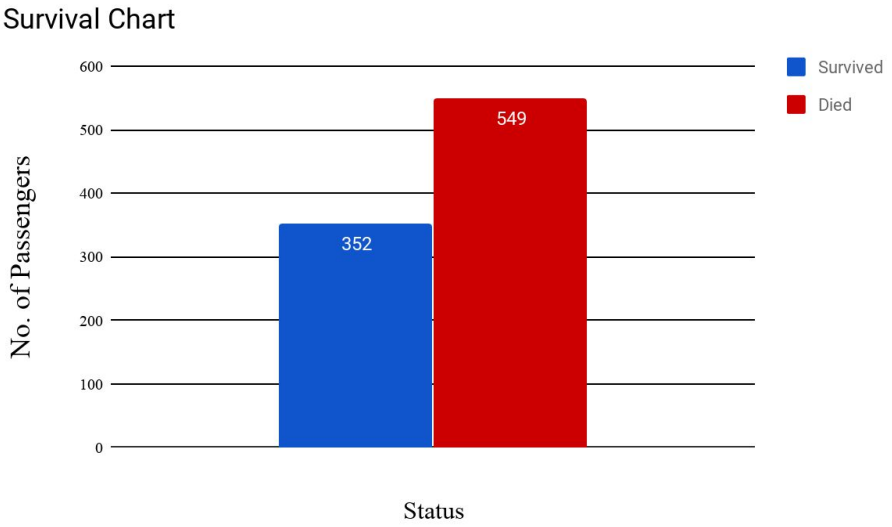
This makes our Model simpler and faster and reduces the unwanted and unnecessary processing.

Dataset

- The given dataset is in the form of .csv file which stands for comma separated file.
- The dataset consists of 891 rows and 12 columns.
- The dataset contains 12 data columns containing values for ***PassengerId***(ID of the Passenger), ***Survived***(Passenger Survived or Died), ***Pclass***(Travelling Class of the Passenger), ***Name***, ***Sex***(Gender of the Passenger), ***Age***, ***SibSp***(Siblings/Spouse), ***Parch*** (Parent/Child), ***Ticket***(Ticket no.), ***Fare***, ***Cabin***(Cabin no. if taken),***Embarked*** (Place of coming aboard).

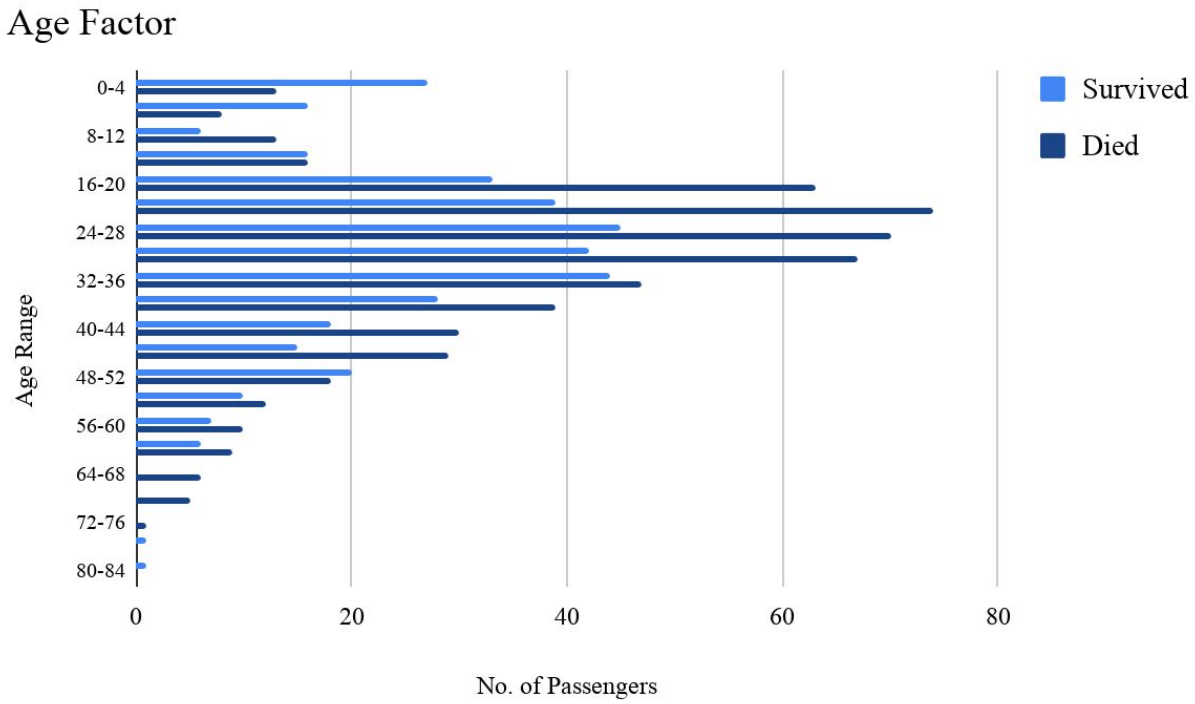
Visualization - Finding the Dependency of Parameters on Survival

Survived : From the total dataset of 891 people the no of people survived are



Visualization - Finding the Dependency of Parameters on Survival

1. Age :

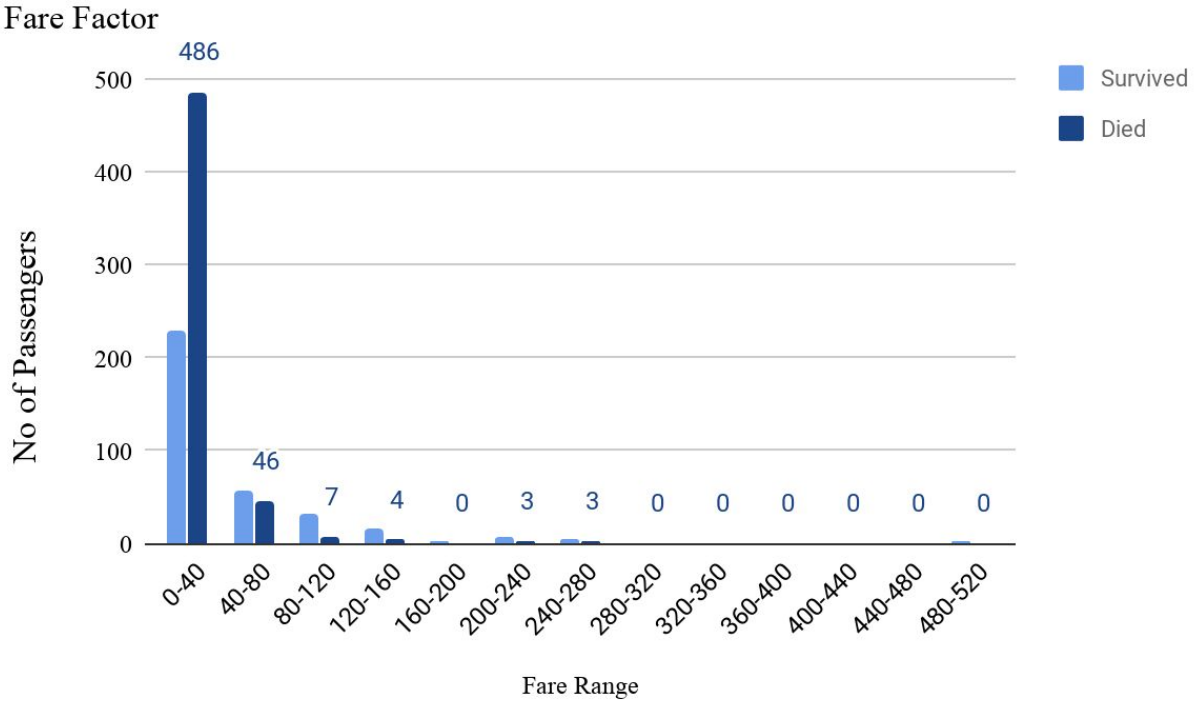


1. According to the graph the age range bw the 16-44 has less survival ratio than the other ranges.

2. The old persons belonging to age group 72-84 also survived as compared to the younger persons in the age range 16-44.

3. The childrens below 16 years of age also has good survival ratio.

2. Fare:

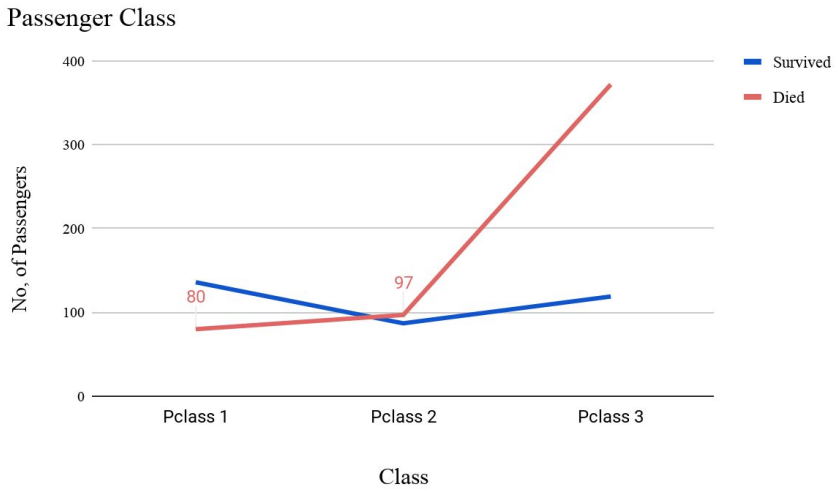


1. According to the graph the passengers who paid fare bw 0-40 the least fare has approximately 50% chance of survival.

2. The passengers with higher fares has more chances of survival.

3. Pclass (Passenger Class)

Pclass	Survived	Died	Percentage (Survival)
1	136	80	0.6296
2	87	97	0.4728
3	119	372	0.2424



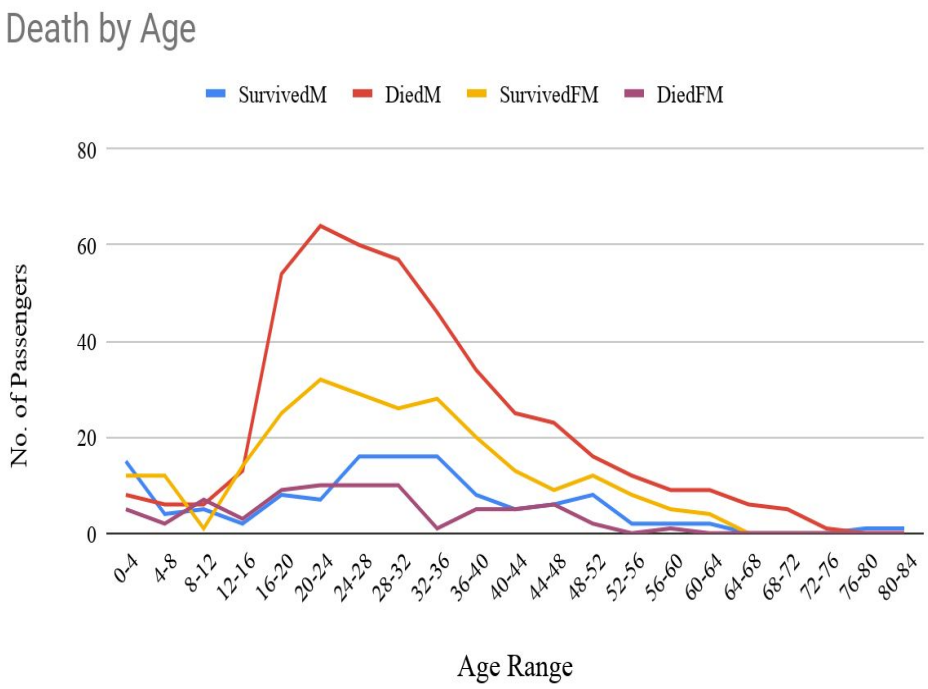
1. According to the table the passengers who were travelling in 3rd class were most hit by the disaster and their percentage of survival was too low.

2. The passengers belonging to 1st class has the highest probability of survival.

3. As like the fare chart the people with high fare must be belonging to the 1st class hence the data in the fare and the Pclass column is dependent on each other.

4.But Pclass can’t be left cause the percentage of survival in higher fare and percentage of survival in 1st class dosen’t match.

4. Sex Factor



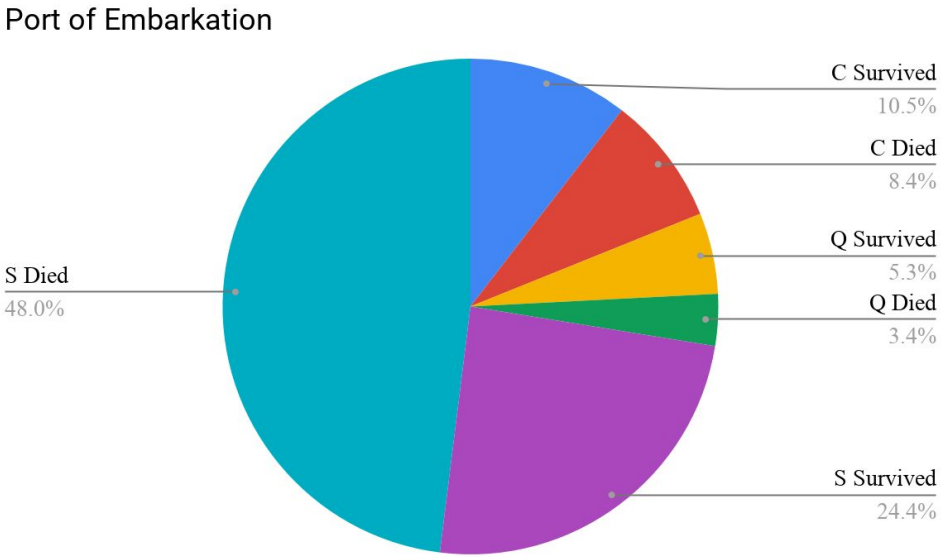
1. According to the table most of the females aboard were saved.

2. The males has survival ratio of about 19% which is very low.

3. Acc. to the graph most of the saved males were either children or older persons.

Sex	Survived	Died	Percentage (Survival)
Female	233	81	0.7420
Male	109	468	0.1889

5. Embarked



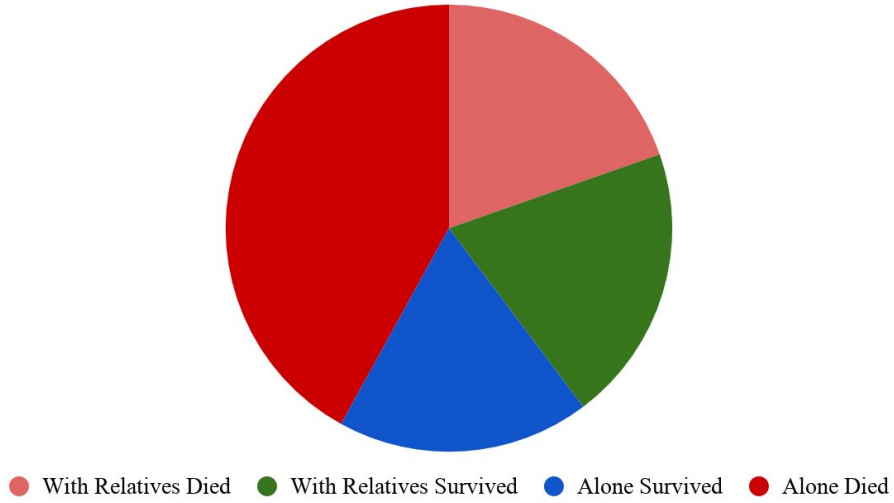
1. According to the table the people who embarked from Southampton has less survival % and most survival % from Cherbourg.

2. Approximately half of the pie graph are the people which died after embarking from Southampton.

Embarked	Survived	Died	Percentage (Survival)
C(Cherbourg)	93	75	0.5535
Q(Queenstown)	30	47	0.3896
S(Southampton)	217	427	0.337

5. Relatives (SibSp,Parch)

Relatives Factor



1. The persons with their family or relatives onboard have more percentage of survival than the persons alone.

Relatives Onboard	Survived	Died	Percentage (Survival)
Yes	179	175	0.5056
No	163	374	0.3035

Data Preprocessing

In this process the data is made ready to be given to a mathematical model so that it can find dependencies in the dataset and can make prediction.

The data fed to a model should have all values as numeric data type as computer can not understand anything other then binary.

The columns chosen for models are :- Age, Fare, Relatives, Embarked, Pclass, Sex

The Relative column was created having values 0 or 1 depending upon the values of Parch and SibSp.

Data Preprocessing

The data contains many null values which makes it unfit for model the Age had the highest null values of 177 and embarked had 2 null values. The Age null values were filled with the average Age and the embarked was filled with most dominant S category.

Age column was scaled bw 0-1 using the StandardScaler as other values are only 0 and 1 and age is bw 0-80 so it will have more weight in the model and induces parity.

Same way the Fare was scaled bw 0-1 using StandardScaler.

As the Sex and Pclass columns contains character values like C, Q, S, female and male so they were encode into five columns using get_dummies function.

Creating A Model

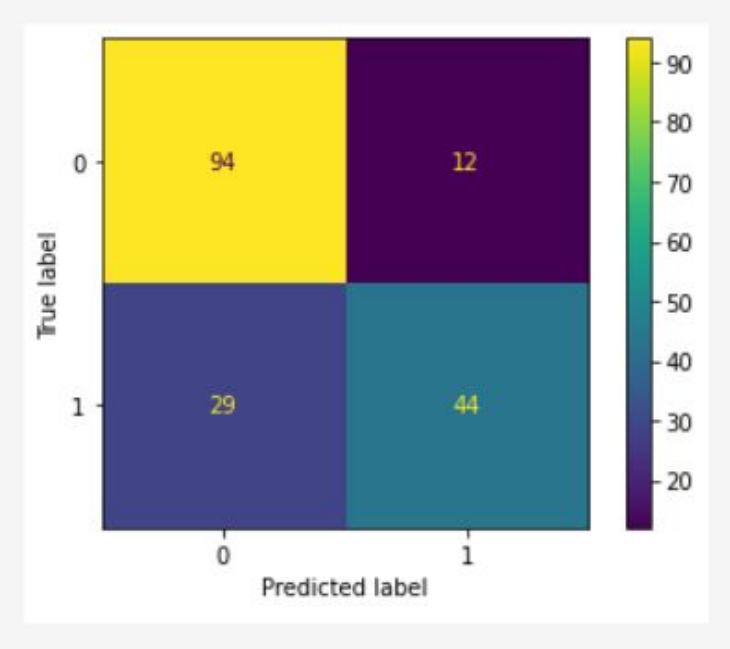
As the data is to be classified into Survived and Died only hence the classifiers are used for the task.

Three Classifiers which are used are:

- KNeighbors Classifier
- Random Forest Classifier
- Logistic Regression
- SVC

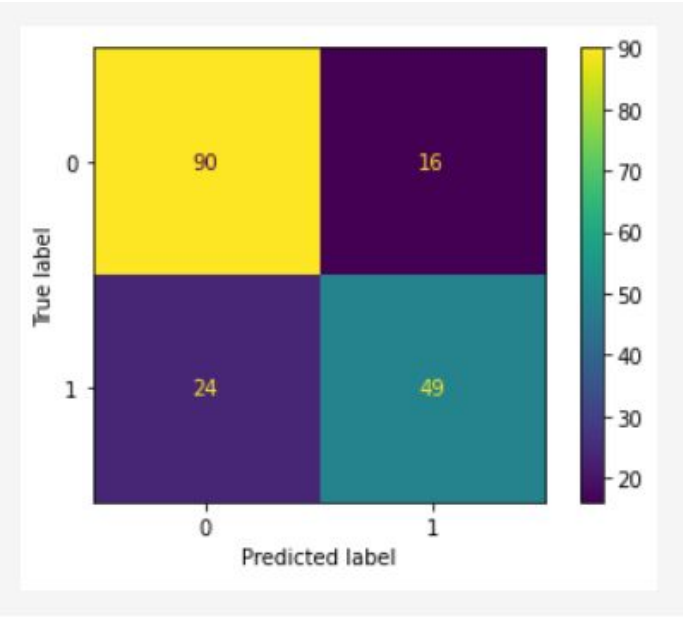
Random Forest Classifier

Accuracy	0.7709



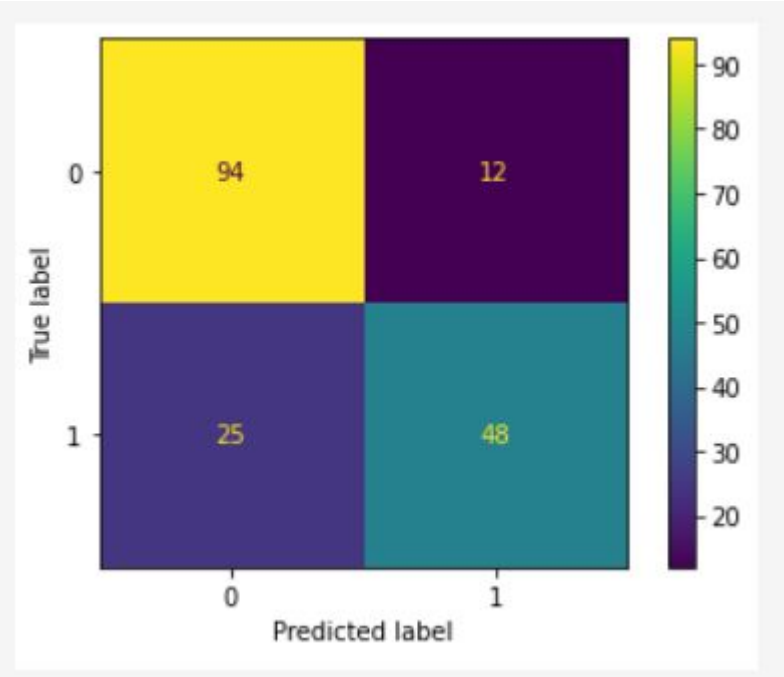
Logistic Regression

Accuracy	0.7765



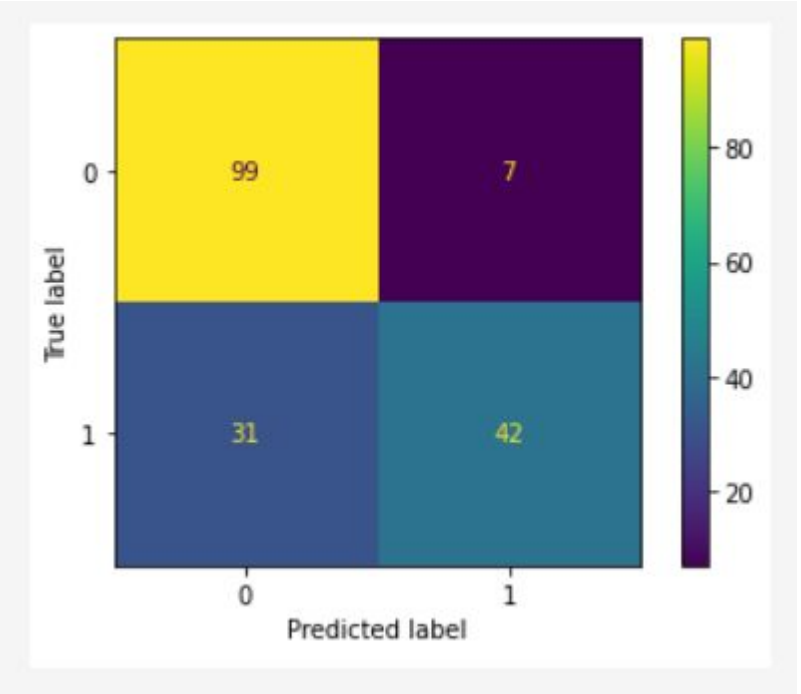
SVC

Accuracy	0.8011



KNeighbors Classifier

Accuracy	0.7877



Thank You