**Project  
 Statistical Learning with R (STAT 6620)  
 Prof. Eric A Suess  
  
  
  
  
  
  
  
  
  
  
 Submitted By  
 Priyanka Ramesh Nayak  
 Bd5787**

# Project Proposal

The data set is Titanic Dataset. The dataset is collected from kaggel <https://www.kaggle.com/c/titanic/data> The data has been split into two groups: . training set (train.csv) . test set (test.csv) The training set should be used to build your machine learning models. For the training set, we provide the outcome (also known as the “ground truth”) for each passenger. The test set should be used to see how well your model performs on unseen data. For the test set, we do not provide the ground truth for each passenger. We were given 891 passenger samples for our training set and their associated labels of whether or not the passenger survived. For each passenger, we were given his/her passenger class, name, sex, age, number ofsiblings/spouses aboard, number of parents/children aboard, ticket number, fare, cabin embarked, and port of embarkation. For the test data, we had 418 samples in the same format. We can use Naïve Bayes Classification algorithm to predict whether or not the passengers survived the sinking of the Titanic depending on the different Class,Sex and Age. Bayes’ theorem can be used to make prediction based on prior knowledge and current evidence. We can also use logistic Regression on survival Survival 0 = No, 1 = Yes. In this project i have used Naive Baeys Classification algorithm to predict the surival of passengers in the sinking of Titanic dependng on Class,Sex and Age.

# Step 1: Collecting Data

The Titanic dataset was collected from kaggel by www.kaggle.com, Using data provided our goal is to apply machine-learning techniques to successfully predict which passengers survived the sinking of the Titanic. Features like ticket price, age, sex, and class will be used to make the predictions. We take several approaches to this problem in order to compare and contrast the different machine learning techniques. In this project i have used Naive Baeys algorithm to predict the survival. NaÃ¯ve Bayes classification is a kind of simple probabilistic classification methods based on Bayes’ theorem with the assumption of independence between features. The model is trained on training dataset to make predictions by predict() function. This article introduces two functions naiveBayes() and train() for the performance of NaÃ¯ve Bayes classification.

# Step 2: Exploring and preparing the data

The first step in constructing the classifier is to import the data and saving it.

library(e1071)

## Warning: package 'e1071' was built under R version 3.4.3

Titanictrain <- read.csv("train.csv", stringsAsFactors = FALSE)  
Titanictest<-read.csv("test.csv", stringsAsFactors = FALSE)

The str() function of test data shows that the data has a total of 418 observations with 11 variables.

str(Titanictest)

## 'data.frame': 418 obs. of 11 variables:  
## $ PassengerId: int 892 893 894 895 896 897 898 899 900 901 ...  
## $ Pclass : int 3 3 2 3 3 3 3 2 3 3 ...  
## $ Name : chr "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...  
## $ Sex : chr "male" "female" "male" "male" ...  
## $ Age : num 34.5 47 62 27 22 14 30 26 18 21 ...  
## $ SibSp : int 0 1 0 0 1 0 0 1 0 2 ...  
## $ Parch : int 0 0 0 0 1 0 0 1 0 0 ...  
## $ Ticket : chr "330911" "363272" "240276" "315154" ...  
## $ Fare : num 7.83 7 9.69 8.66 12.29 ...  
## $ Cabin : chr "" "" "" "" ...  
## $ Embarked : chr "Q" "S" "Q" "S" ...

The str() function of train data shows that the data has a total of 891 observations with 12 variables.

str(Titanictrain)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr "" "C85" "" "C123" ...  
## $ Embarked : chr "S" "C" "S" "S" ...

titanicdata<-naiveBayes(as.factor(Survived)~., Titanictrain)  
str(titanicdata)

## List of 4  
## $ apriori: 'table' int [1:2(1d)] 549 342  
## ..- attr(\*, "dimnames")=List of 1  
## .. ..$ Y: chr [1:2] "0" "1"  
## $ tables :List of 11  
## ..$ PassengerId: num [1:2, 1:2] 447 444 261 252  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ PassengerId: NULL  
## ..$ Pclass : num [1:2, 1:2] 2.532 1.95 0.736 0.863  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Pclass: NULL  
## ..$ Name : table [1:2, 1:891] 0.00182 0 0.00182 0 0 ...  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Name: chr [1:891] "Abbing, Mr. Anthony" "Abbott, Mr. Rossmore Edward" "Abbott, Mrs. Stanton (Rosa Hunt)" "Abelson, Mr. Samuel" ...  
## ..$ Sex : table [1:2, 1:2] 0.148 0.681 0.852 0.319  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Sex: chr [1:2] "female" "male"  
## ..$ Age : num [1:2, 1:2] 30.6 28.3 14.2 15  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Age: NULL  
## ..$ SibSp : num [1:2, 1:2] 0.554 0.474 1.288 0.709  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ SibSp: NULL  
## ..$ Parch : num [1:2, 1:2] 0.33 0.465 0.823 0.772  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Parch: NULL  
## ..$ Ticket : table [1:2, 1:681] 0 0.00877 0.00182 0.00585 0.00364 ...  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Ticket: chr [1:681] "110152" "110413" "110465" "110564" ...  
## ..$ Fare : num [1:2, 1:2] 22.1 48.4 31.4 66.6  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Fare: NULL  
## ..$ Cabin : table [1:2, 1:148] 0.87614 0.60234 0.00182 0 0.00182 ...  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Cabin: chr [1:148] "" "A10" "A14" "A16" ...  
## ..$ Embarked : table [1:2, 1:4] 0 0.00585 0.13661 0.27193 0.08561 ...  
## .. ..- attr(\*, "dimnames")=List of 2  
## .. .. ..$ Y : chr [1:2] "0" "1"  
## .. .. ..$ Embarked: chr [1:4] "" "C" "Q" "S"  
## $ levels : chr [1:2] "0" "1"  
## $ call : language naiveBayes.default(x = X, y = Y, laplace = laplace)  
## - attr(\*, "class")= chr "naiveBayes"

###Relabelling some categorical variables  
####Survived  
Titanictrain$Survived[Titanictrain$Survived==0]="Not survived"  
Titanictrain$Survived[Titanictrain$Survived==1]="Survived"  
  
table(Titanictrain$Survived)

##   
## Not survived Survived   
## 549 342

# Number of survivals and non-survivals by Gender

table(Titanictrain$Survived,Titanictrain$Sex)

##   
## female male  
## Not survived 81 468  
## Survived 233 109

### Number of survivals and non-survivals by Class

table(Titanictrain$Survived,Titanictrain$Pclass)

##   
## 1 2 3  
## Not survived 80 97 372  
## Survived 136 87 119

table(Titanictrain$Survived,Titanictrain$Embarked)

##   
## C Q S  
## Not survived 0 75 47 427  
## Survived 2 93 30 217

table(Titanictrain$Survived,Titanictrain$Age)

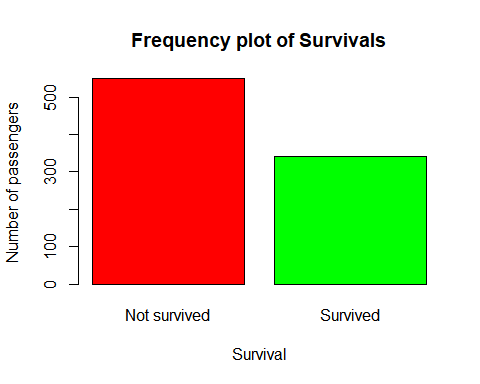
##   
## 0.42 0.67 0.75 0.83 0.92 1 2 3 4 5 6 7 8 9 10 11  
## Not survived 0 0 0 0 0 2 7 1 3 0 1 2 2 6 2 3  
## Survived 1 1 2 2 1 5 3 5 7 4 2 1 2 2 0 1  
##   
## 12 13 14 14.5 15 16 17 18 19 20 20.5 21 22 23 23.5 24 24.5  
## Not survived 0 0 3 1 1 11 7 17 16 12 1 19 16 10 1 15 1  
## Survived 1 2 3 0 4 6 6 9 9 3 0 5 11 5 0 15 0  
##   
## 25 26 27 28 28.5 29 30 30.5 31 32 32.5 33 34 34.5 35 36  
## Not survived 17 12 7 18 2 12 15 2 9 9 1 9 9 1 7 11  
## Survived 6 6 11 7 0 8 10 0 8 9 1 6 6 0 11 11  
##   
## 36.5 37 38 39 40 40.5 41 42 43 44 45 45.5 46 47 48 49 50 51  
## Not survived 1 5 6 9 7 2 4 7 4 6 7 2 3 8 3 2 5 5  
## Survived 0 1 5 5 6 0 2 6 1 3 5 0 0 1 6 4 5 2  
##   
## 52 53 54 55 55.5 56 57 58 59 60 61 62 63 64 65 66 70 70.5  
## Not survived 3 0 5 1 1 2 2 2 2 2 3 2 0 2 3 1 2 1  
## Survived 3 1 3 1 0 2 0 3 0 2 0 2 2 0 0 0 0 0  
##   
## 71 74 80  
## Not survived 2 1 0  
## Survived 0 0 1

# visulization of the data

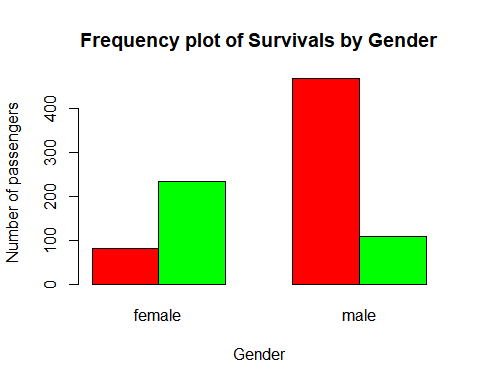
# Frequency plot of survivals and non-survivals

A Bar Plot is a way to visualize Frequency plot of survivals and non-survivals

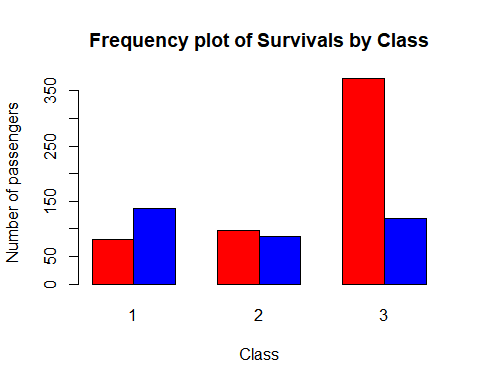
barplot(table(Titanictrain$Survived),col=c("red","green"),main="Frequency plot of Survivals",xlab="Survival",ylab="Number of passengers")

 #Frequency plot of survivals and non-survivals by Gender

barplot(table(Titanictrain$Survived,Titanictrain$Sex),col=c("red","green"),beside=TRUE,main="Frequency plot of Survivals by Gender",xlab="Gender",ylab="Number of passengers")

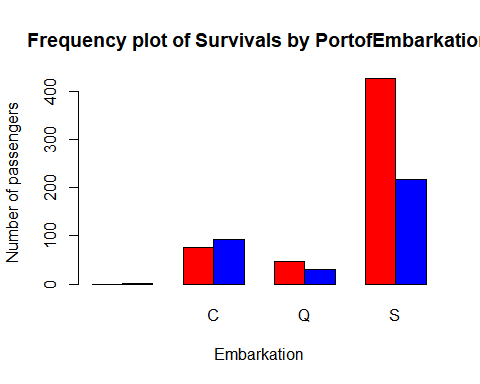
 #Frequency plot of survivals and non-survivals by Class

barplot(table(Titanictrain$Survived,Titanictrain$Pclass),col=c("red","blue"),beside=TRUE, main="Frequency plot of Survivals by Class",xlab="Class",ylab="Number of passengers")



# Frequency plot of survivals and non-survivals by Port of Embarkation

barplot(table(Titanictrain$Survived,Titanictrain$Embarked),col=c("red","blue"), beside=TRUE, main="Frequency plot of Survivals by PortofEmbarkation",xlab="Embarkation",ylab="Number of passengers")



# Data preparation:

# Data Cleaning and Preparation

dt=data.frame(as.factor(Titanictrain$Survived),Titanictrain[,-2])  
dt1=dt[,-2] ##Removing Passenger ID###  
dt2=dt1[,-3] ###Removing Passenger Name###  
dt3=dt2[,-7] ####Removing Ticket#####  
clean\_data=dt3[,-8] ####Removing Cabin####  
varnames=names(clean\_data)  
varnames=c("Survived",varnames[-1])  
colnames(clean\_data)=varnames  
print("The following are the predictor variables used:")

## [1] "The following are the predictor variables used:"

names(clean\_data[,2:ncol(clean\_data)])

## [1] "Pclass" "Sex" "Age" "SibSp" "Parch" "Fare"   
## [7] "Embarked"

# Model Fitting and Evaluation

n=nrow(clean\_data)

# creating training and test datasets

With our data prepared for analysis, we now need to split the data into training and test datasets.Titanic dataset has traing and test datset but the test dataset doesnot have the surivalid.so we have to devide Testdata set as Traning dataset and Validation dataset.We’ll divide the data into two portions: 75 percent for training and 25 percent for validation data testing.

# creating training and test datasets  
t\_train <- clean\_data[1:668, ]  
t\_test <- clean\_data[669:891, ]

## Step 3: Training a model on the data —-

we have transformed the raw titanic training data into a format that can be represented by a statistical model, it is time to apply the Naive Bayes algorithm. The algorithm will predict the survival adn non survival of passengers in the Titanic ship. The Naive Bayes implementation we will employ is in the e1071 package. This package was developed in the statistics department of the Vienna University ofTechnology (TU Wien)

library(e1071)  
Titanic\_classifier <- naiveBayes(Survived~.,data=t\_train)

## Step 4: Evaluating model performance —-

To evaluate the Titanic classifier, we need to test its predictions train data. The classifier that we trained has been named sms\_classifier.We will use this classifier to generate predictions and then compare the predicted values to the true values. The predict() function is used to make the predictions. We will store these in a vector named Titanic\_test\_pred We will simply supply the function with the names of our classifier and test dataset, as shown.

Titanic\_test\_pred <- predict(Titanic\_classifier,t\_test[,-1])

## Warning in data.matrix(newdata): NAs introduced by coercion  
  
## Warning in data.matrix(newdata): NAs introduced by coercion

head(data.frame(Titanic\_test\_pred,t\_test))

## Titanic\_test\_pred Survived Pclass Sex Age SibSp Parch Fare  
## 669 Not survived Not survived 3 male 43 0 0 8.05  
## 670 Survived Survived 1 female NA 1 0 52.00  
## 671 Not survived Survived 2 female 40 1 1 39.00  
## 672 Survived Not survived 1 male 31 1 0 52.00  
## 673 Not survived Not survived 2 male 70 0 0 10.50  
## 674 Not survived Survived 2 male 31 0 0 13.00  
## Embarked  
## 669 S  
## 670 S  
## 671 S  
## 672 S  
## 673 S  
## 674 S

To compare the predictions to the true values, we’ll use the CrossTable() function in the gmodels package, which we used previously. This time, we’ll add some additional parameters to eliminate unnecessary cell proportions and use the dnn parameter (dimension names) to relabel the rows and columns, as shown in the following code:

library(gmodels)

## Warning: package 'gmodels' was built under R version 3.4.4

CrossTable(Titanic\_test\_pred, t\_test$Survived,prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,  
 dnn = c('predicted', 'actual'))

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 223   
##   
##   
## | actual   
## predicted | Not survived | Survived | Row Total |   
## -------------|--------------|--------------|--------------|  
## Not survived | 125 | 40 | 165 |   
## | 0.887 | 0.488 | |   
## -------------|--------------|--------------|--------------|  
## Survived | 16 | 42 | 58 |   
## | 0.113 | 0.512 | |   
## -------------|--------------|--------------|--------------|  
## Column Total | 141 | 82 | 223 |   
## | 0.632 | 0.368 | |   
## -------------|--------------|--------------|--------------|  
##   
##

The CrossTable above shows that 56 passengers were incorrectly classified in which 16 out of 141 not survived were misidentified as survived, and 40 out of 82 Survived were misdentified as nonsurvived.

The accuarcy of the model can be calculated as

# Calculate the accuracy

error <- mean(t\_test$Survived != Titanic\_test\_pred) # Misclassification error  
  
paste('Accuracy',round(1-error,4))

## [1] "Accuracy 0.7489"

# Step 5: Improving model performance

We can try to improve the model by setting value for the laplace estimator.We’ll build a Naive Bayes model as done earlier, but this time set laplace = 1.

Titanic\_classifier2 <- naiveBayes(Survived~.,data=t\_train,laplace = 1)  
  
  
Titanic\_test\_pred2 <- predict(Titanic\_classifier2,t\_test[,-1])

## Warning in data.matrix(newdata): NAs introduced by coercion  
  
## Warning in data.matrix(newdata): NAs introduced by coercion

head(data.frame(Titanic\_test\_pred2,t\_test))

## Titanic\_test\_pred2 Survived Pclass Sex Age SibSp Parch Fare  
## 669 Not survived Not survived 3 male 43 0 0 8.05  
## 670 Survived Survived 1 female NA 1 0 52.00  
## 671 Not survived Survived 2 female 40 1 1 39.00  
## 672 Survived Not survived 1 male 31 1 0 52.00  
## 673 Not survived Not survived 2 male 70 0 0 10.50  
## 674 Not survived Survived 2 male 31 0 0 13.00  
## Embarked  
## 669 S  
## 670 S  
## 671 S  
## 672 S  
## 673 S  
## 674 S

library(gmodels)  
CrossTable(Titanic\_test\_pred2,t\_test$Survived, prop.chisq = FALSE, prop.t = FALSE, prop.r = FALSE,  
 dnn = c('predicted', 'actual'))

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 223   
##   
##   
## | actual   
## predicted | Not survived | Survived | Row Total |   
## -------------|--------------|--------------|--------------|  
## Not survived | 125 | 40 | 165 |   
## | 0.887 | 0.488 | |   
## -------------|--------------|--------------|--------------|  
## Survived | 16 | 42 | 58 |   
## | 0.113 | 0.512 | |   
## -------------|--------------|--------------|--------------|  
## Column Total | 141 | 82 | 223 |   
## | 0.632 | 0.368 | |   
## -------------|--------------|--------------|--------------|  
##   
##

error <- mean(t\_test$Survived != Titanic\_test\_pred2) # Misclassification error  
  
paste('Accuracy',round(1-error,4))

## [1] "Accuracy 0.7489"

By adding laplace estimator it didnt improve the model performance.We got the same accuracy as before.

# Prediction using the test dataset

dl1=Titanictest[,-1] ##Removing Passenger ID###  
dl2=dl1[,-2] ###Removing Passenger Name###  
dl3=dl2[,-6] ####Removing Ticket#####  
test=dl3[,-7] ####Removing Cabin####  
  
  
test$Embarked=as.character(test$Embarked)  
test$Embarked=as.factor(test$Embarked)  
yhat=predict(Titanic\_classifier,test)

## Warning in data.matrix(newdata): NAs introduced by coercion

pred1=data.frame(yhat,Titanictest)  
head(pred1,n=20)

## yhat PassengerId Pclass  
## 1 Not survived 892 3  
## 2 Not survived 893 3  
## 3 Not survived 894 2  
## 4 Not survived 895 3  
## 5 Not survived 896 3  
## 6 Not survived 897 3  
## 7 Not survived 898 3  
## 8 Not survived 899 2  
## 9 Survived 900 3  
## 10 Not survived 901 3  
## 11 Not survived 902 3  
## 12 Survived 903 1  
## 13 Survived 904 1  
## 14 Not survived 905 2  
## 15 Survived 906 1  
## 16 Survived 907 2  
## 17 Not survived 908 2  
## 18 Survived 909 3  
## 19 Not survived 910 3  
## 20 Not survived 911 3  
## Name Sex Age  
## 1 Kelly, Mr. James male 34.5  
## 2 Wilkes, Mrs. James (Ellen Needs) female 47.0  
## 3 Myles, Mr. Thomas Francis male 62.0  
## 4 Wirz, Mr. Albert male 27.0  
## 5 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0  
## 6 Svensson, Mr. Johan Cervin male 14.0  
## 7 Connolly, Miss. Kate female 30.0  
## 8 Caldwell, Mr. Albert Francis male 26.0  
## 9 Abrahim, Mrs. Joseph (Sophie Halaut Easu) female 18.0  
## 10 Davies, Mr. John Samuel male 21.0  
## 11 Ilieff, Mr. Ylio male NA  
## 12 Jones, Mr. Charles Cresson male 46.0  
## 13 Snyder, Mrs. John Pillsbury (Nelle Stevenson) female 23.0  
## 14 Howard, Mr. Benjamin male 63.0  
## 15 Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood) female 47.0  
## 16 del Carlo, Mrs. Sebastiano (Argenia Genovesi) female 24.0  
## 17 Keane, Mr. Daniel male 35.0  
## 18 Assaf, Mr. Gerios male 21.0  
## 19 Ilmakangas, Miss. Ida Livija female 27.0  
## 20 Assaf Khalil, Mrs. Mariana (Miriam")" female 45.0  
## SibSp Parch Ticket Fare Cabin Embarked  
## 1 0 0 330911 7.8292 Q  
## 2 1 0 363272 7.0000 S  
## 3 0 0 240276 9.6875 Q  
## 4 0 0 315154 8.6625 S  
## 5 1 1 3101298 12.2875 S  
## 6 0 0 7538 9.2250 S  
## 7 0 0 330972 7.6292 Q  
## 8 1 1 248738 29.0000 S  
## 9 0 0 2657 7.2292 C  
## 10 2 0 A/4 48871 24.1500 S  
## 11 0 0 349220 7.8958 S  
## 12 0 0 694 26.0000 S  
## 13 1 0 21228 82.2667 B45 S  
## 14 1 0 24065 26.0000 S  
## 15 1 0 W.E.P. 5734 61.1750 E31 S  
## 16 1 0 SC/PARIS 2167 27.7208 C  
## 17 0 0 233734 12.3500 Q  
## 18 0 0 2692 7.2250 C  
## 19 1 0 STON/O2. 3101270 7.9250 S  
## 20 0 0 2696 7.2250 C

# Conclusion

In this project i have used Naive Baeys Classification algorithm to predict which passengers survived in the sinking of the Titanic.Features like ticket price, age, sex, and class will be used to make the predictions.We got the accuracy as 75%.To improve the Model performance, used laplace estimator there were no significant differences in accuracy.The accuracy was same as before.