Titanic Survival Analysis using Decision Tree

Rishabh Lavangad

23 November 2018

1. Import the data set in R, create a variable called titanic & store the data in it.

```
titanic <- read.csv("Titanic.csv")
head(titanic)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                            Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                             Heikkinen, Miss. Laina female  26     0
## 4       Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                           Allen, Mr. William Henry   male  35     0
## 6                                   Moran, Mr. James   male  NA     0
##   Parch            Ticket    Fare Cabin Embarked
## 1     0         A/5 21171  7.2500              S
## 2     0          PC 17599 71.2833   C85        C
## 3     0 STON/O2. 3101282  7.9250              S
## 4     0            113803 53.1000  C123        S
## 5     0            373450  8.0500              S
## 6     0            330877  8.4583              Q
```

2. Print the structure of dataset.

```
str(titanic)
```

```
## '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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ 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     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

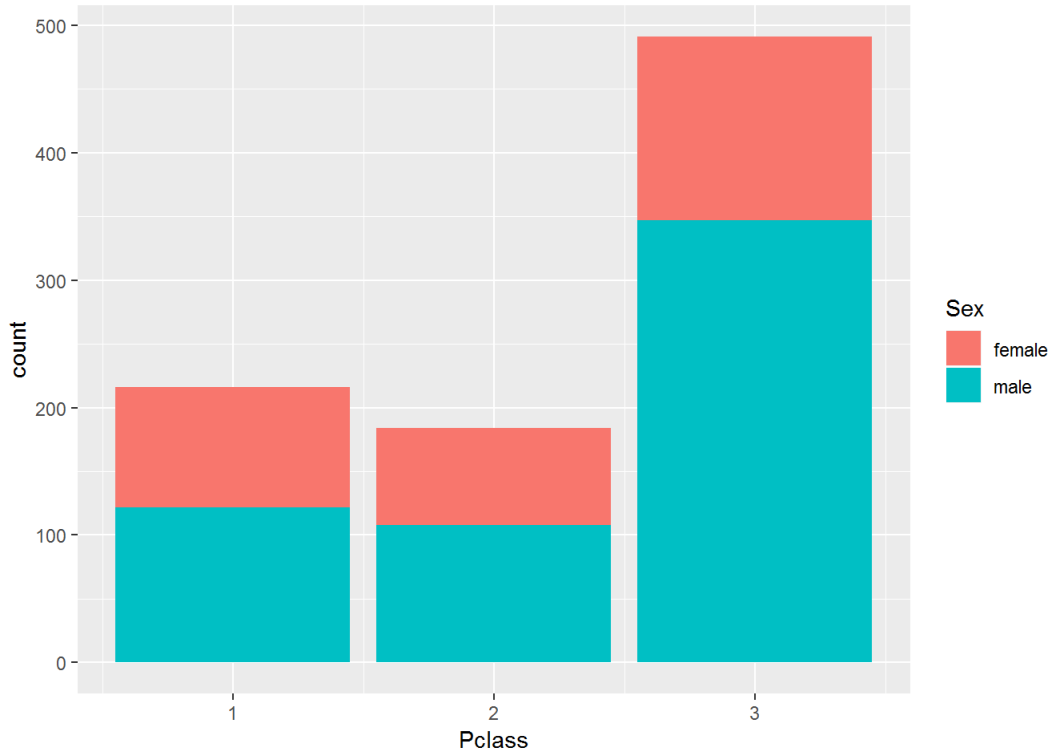
3. Print last 6 rows of dataset.

```
tail(titanic,6)
```

```
##     PassengerId Survived Pclass                                     Name
## 886         886        0      3      Rice, Mrs. William (Margaret Norton)
## 887         887        0      2                     Montvila, Rev. Juozas
## 888         888        1      1              Graham, Miss. Margaret Edith
## 889         889        0      3  Johnston, Miss. Catherine Helen "Carrie"
## 890         890        1      1                     Behr, Mr. Karl Howell
## 891         891        0      3                       Dooley, Mr. Patrick
##        Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
## 886 female  39     0     5     382652 29.125              Q
## 887   male  27     0     0     211536 13.000              S
## 888 female  19     0     0     112053 30.000   B42        S
## 889 female  NA     1     2 W./C. 6607 23.450              S
## 890   male  26     0     0     111369 30.000  C148        C
## 891   male  32     0     0     370376  7.750              Q
```

4.Create a barplot using ggplot with class on x-axis, frequency on y -axis & sex as the fill, write your observation.

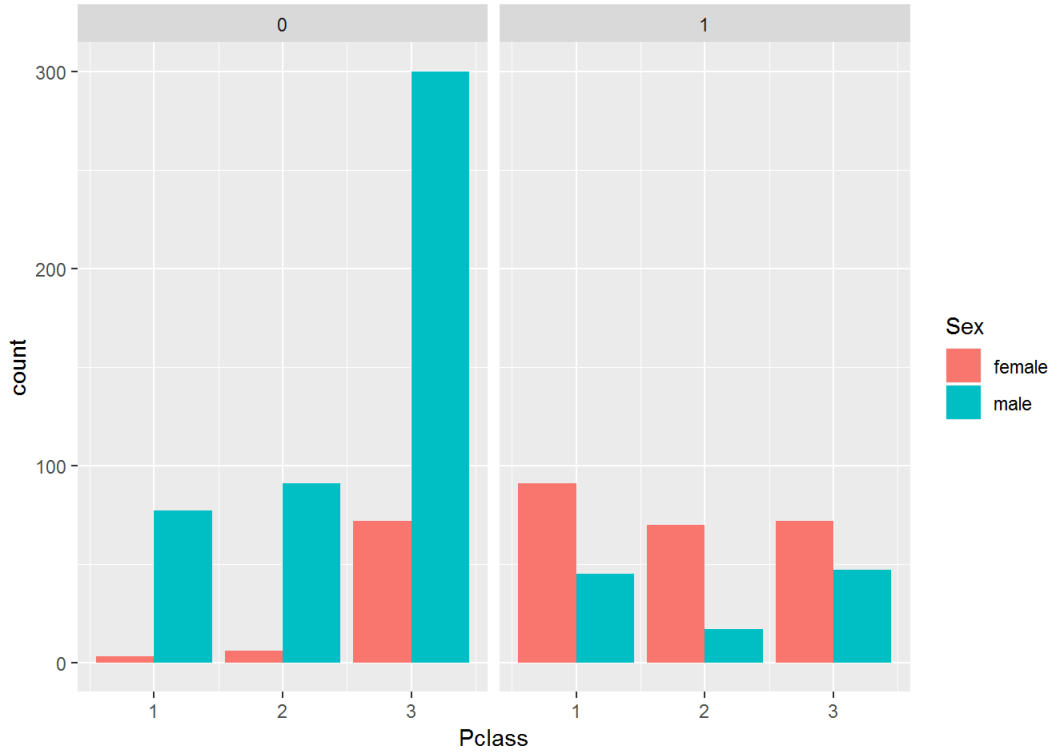
```
library(ggplot2)
ggplot(titanic, aes(x = Pclass, fill = Sex)) + geom_bar()
```



OBSERVATION :- As compared to first two Pclass's,the third Pclass has more number of male and female count.

5. Use ggplot() to estimate your chances of servival from the distibution of sexes within the classes of the ship; Hint: use fact grids.

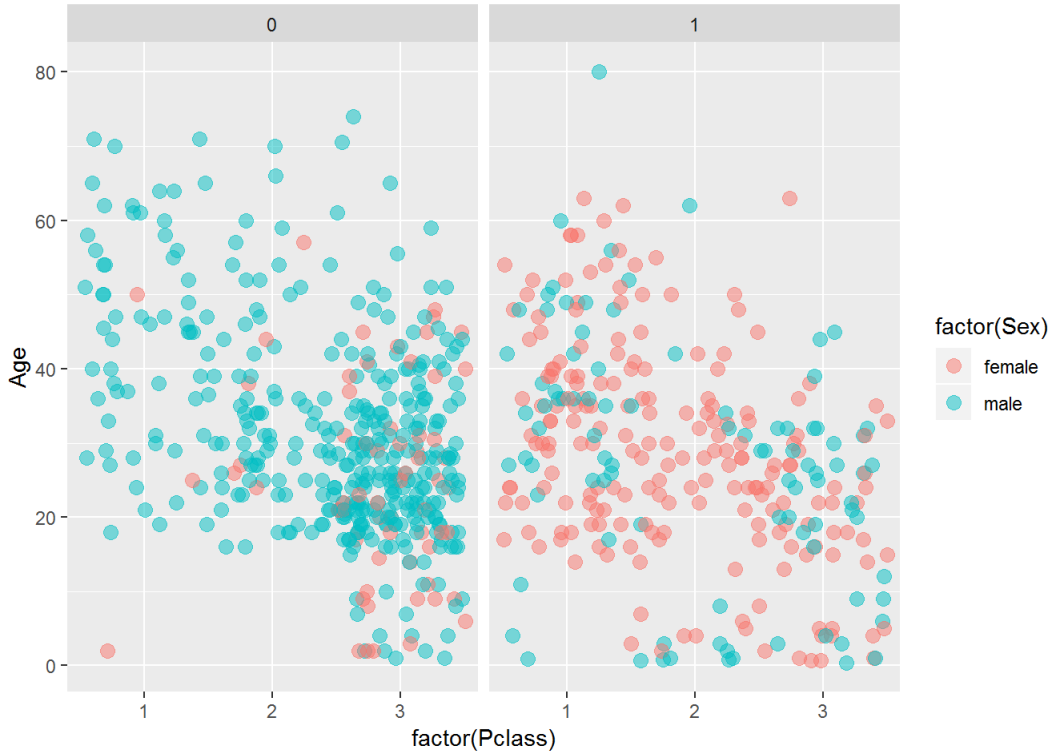
```
ggplot(titanic, aes(x = Pclass, fill = Sex)) + geom_bar(position = "dodge") + facet_grid(~Survived)
```



6. Use ggplot() to estimate your chances of survival based on your age from the distribution of sexes within the class of the ship; Hint: Use the above plot & overlay age using jitter plot .

```
posn.j <- position_jitter(0.5, 0)
ggplot(titanic,aes(x=factor(Pclass),y=Age,col=factor(Sex)))+
        geom_jitter(size=3,alpha=0.5,position=posn.j)+
        facet_grid(". ~ Survived")
```

```
## Warning: Removed 177 rows containing missing values (geom_point).
```



7. Import the training dataset using the following link and print the structure to console "http://50.amazonaws.com/assets.datacamp.com/course/kaggle/train.csv"

```
train_url <- "http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/train.csv"
train <- read.csv(train_url)
str(train)
```

```
## '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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ 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     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

8. Import the testing dataset using the following link & print the structure to console "http://50.amazonaws.com/assets.datacamp.com/course/kaggle/train.csv"

```
test_url <- "http://s3.amazonaws.com/assets.datacamp.com/course/Kaggle/test.csv"
test <- read.csv(test_url)
str(test)
```

```
## '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       : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85 58 5
104 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
##  $ 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     : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270 ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

9. Print the survival rates in absolute numbers for train.

```
table(train$Survived)
```

```
##
##   0   1
## 549 342
```

10. Print th survival rates in proportion for train.

```
prop.table(table(train$Survived))
```

```
##
##         0         1
## 0.6161616 0.3838384
```

11. Print a two way comparision table for sex and survived

```
table(train$Sex, train$Survived)
```

```
##
##            0   1
##   female  81 233
##   male   468 109
```

12. Print a two way comparision for sex and survived, show the proportions row wise i.e for Male & Female

```
prop.table(table(train$Sex, train$Survived),margin =1)
```

```
##
##                  0         1
##   female 0.2579618 0.7420382
##   male   0.8110919 0.1889081
```

13. Create a column called child using the condition age < 18

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
train$Child <- NA
train$Child[train$Age < 18] <- "1"
train$Child[train$Age >= 18] <- "0"
head(train)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                 Name    Sex Age SibSp
## 1                            Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                             Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                           Allen, Mr. William Henry   male  35     0
## 6                                   Moran, Mr. James   male  NA     0
##   Parch            Ticket    Fare Cabin Embarked Child
## 1     0         A/5 21171  7.2500              S     0
## 2     0          PC 17599 71.2833   C85        C     0
## 3     0 STON/O2. 3101282  7.9250              S     0
## 4     0            113803 53.1000  C123        S     0
## 5     0            373450  8.0500              S     0
## 6     0            330877  8.4583              Q  <NA>
```

14. Print a two way comparison table to show the proportion of this child variable with respect to survival

```
prop.table(table(train$Child,train$Survived))
```

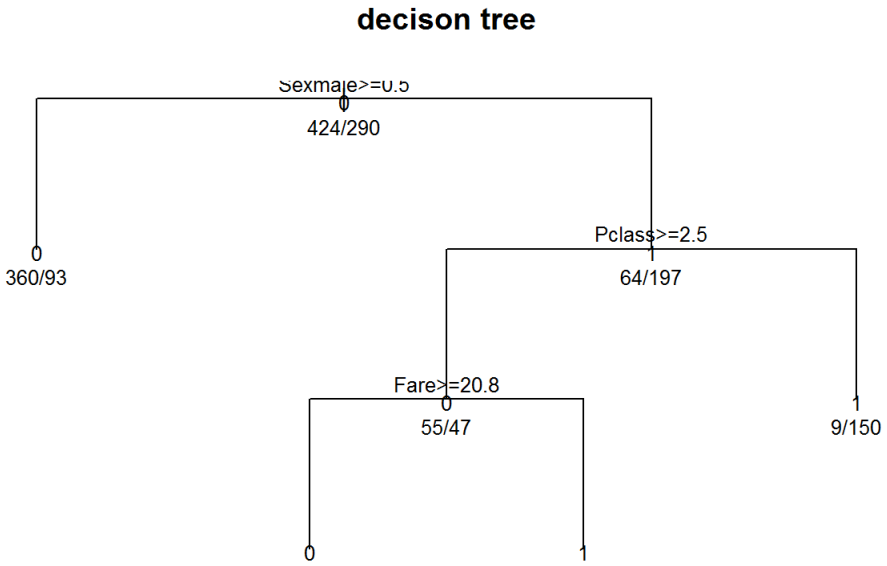
```
##
##              0          1
##   0 0.52100840 0.32072829
##   1 0.07282913 0.08543417
```

15. Using rpart build a decision tree on the training dataset Dependent Vaar : Survived Independent var : Pclass,sex,age,parch,fare,embarked

```
library(caret)
```

```
## Loading required package: lattice
```

```
train$Survived=as.factor(train$Survived)
tree <- train(Survived  ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked, data = train, method = "rpa
rt", na.action = na.exclude)
plot(tree$finalModel,uniform = T,main="decison tree")
text(tree$finalModel,use.n = T,all=T,cex=.8)
```

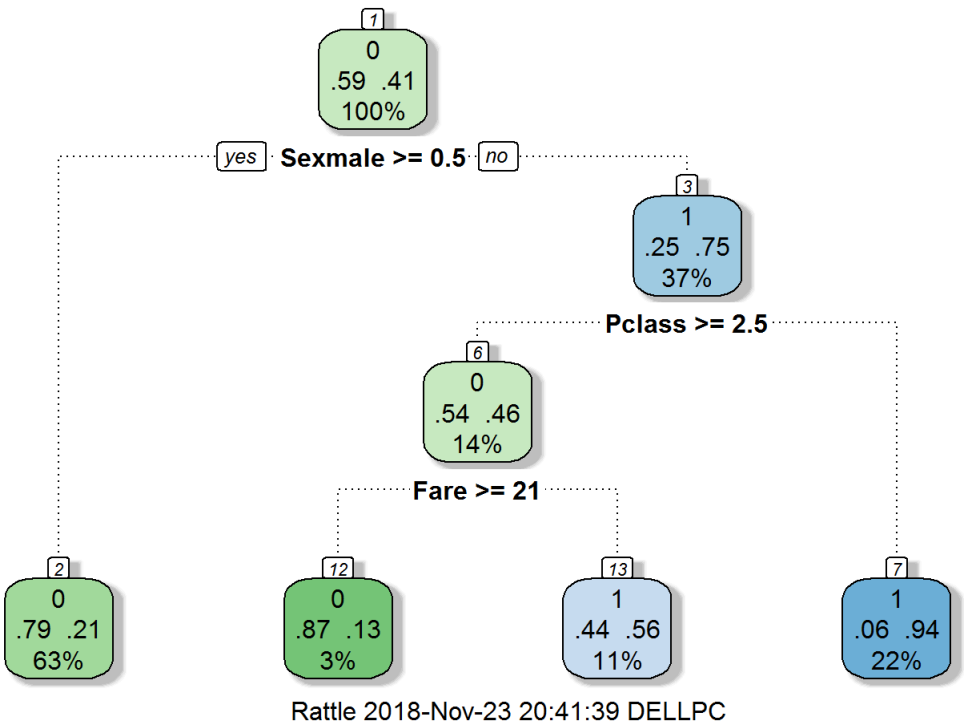


16. Plot your decision tee using fancyrpart plots

```
library(rattle)
```

```
## Rattle: A free graphical interface for data science with R.
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
fancyRpartPlot(tree$finalModel)
```



17. Make predictions on the test set

```
prediction <- predict(tree, test,type = "prob")
head(prediction)
```

```
##          0        1
## 1 0.794702 0.205298
## 2 0.443038 0.556962
## 3 0.794702 0.205298
## 4 0.794702 0.205298
## 5 0.443038 0.556962
## 6 0.794702 0.205298
```