Project1NB.R

RaxyR

2021-03-07

```
#title: "CS 4375 Project 1 Naive Bayes"
#author: "Ramesh Kanakala"
#subtitle: "This is an R script with the purpose of running naive bayes on a
#titanic data set to observe run time and other metrics"

### Logistic Regression
#load the data
ttnc <- read.csv(file = 'titanic_project.csv')
ttnc$pclass <- as.factor(ttnc$pclass)
ttnc$sex <- as.factor(ttnc$sex)
ttnc$survived <- as.factor(ttnc$survived)

#dividing into train/test, putting 75% in train
i <- 1:900
train <- ttnc[i,]
test <- ttnc[-i,]

start <- Sys.time()
#train naive bayes model
library(e1071)
nb1 <- naiveBayes(as.factor(survived)~pclass+sex+age,family = "binomial", data = train)
end <- Sys.time()

#print probabilities from model
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace, family = "binomial")
##
## A-priori probabilities:
## Y
##   0   1
## 0.6 0.4
##
## Conditional probabilities:
##    pclass
## Y          1         2         3
##   0 0.1685185 0.2203704 0.6111111
##   1 0.4166667 0.2638889 0.3194444
```

```
##
##    sex
## Y           0         1
##   0 0.1592593 0.8407407
##   1 0.6944444 0.3055556
##
##    age
## Y       [,1]     [,2]
##   0 30.41682 14.21185
##   1 28.92060 15.09074
```

```
#test on test data
pred <- predict(nb1, newdata=test, type="class")

#print accuracy, sensitivity, and specificity #check spelling
library(caret)
```

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

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))$overall[1]
```

```
## Accuracy
## 0.760274
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))$byClass[1]
```

```
## Sensitivity
##   0.8734177
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))$byClass[2]
```

```
## Specificity
##   0.6268657
```

```
confusionMatrix(as.factor(pred), as.factor(test$survived))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 69 25
##          1 10 42
##
##                Accuracy : 0.7603
##                  95% CI : (0.6827, 0.827)
##     No Information Rate : 0.5411
```

```
##     P-Value [Acc > NIR] : 3.612e-08
##
##                   Kappa : 0.5089
##
##  Mcnemar's Test P-Value : 0.01796
##
##             Sensitivity : 0.8734
##             Specificity : 0.6269
##          Pos Pred Value : 0.7340
##          Neg Pred Value : 0.8077
##              Prevalence : 0.5411
##          Detection Rate : 0.4726
##    Detection Prevalence : 0.6438
##       Balanced Accuracy : 0.7501
##
##        'Positive' Class : 0
##
```

```
#time difference
end - start
```

```
## Time difference of 0.03402996 secs
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

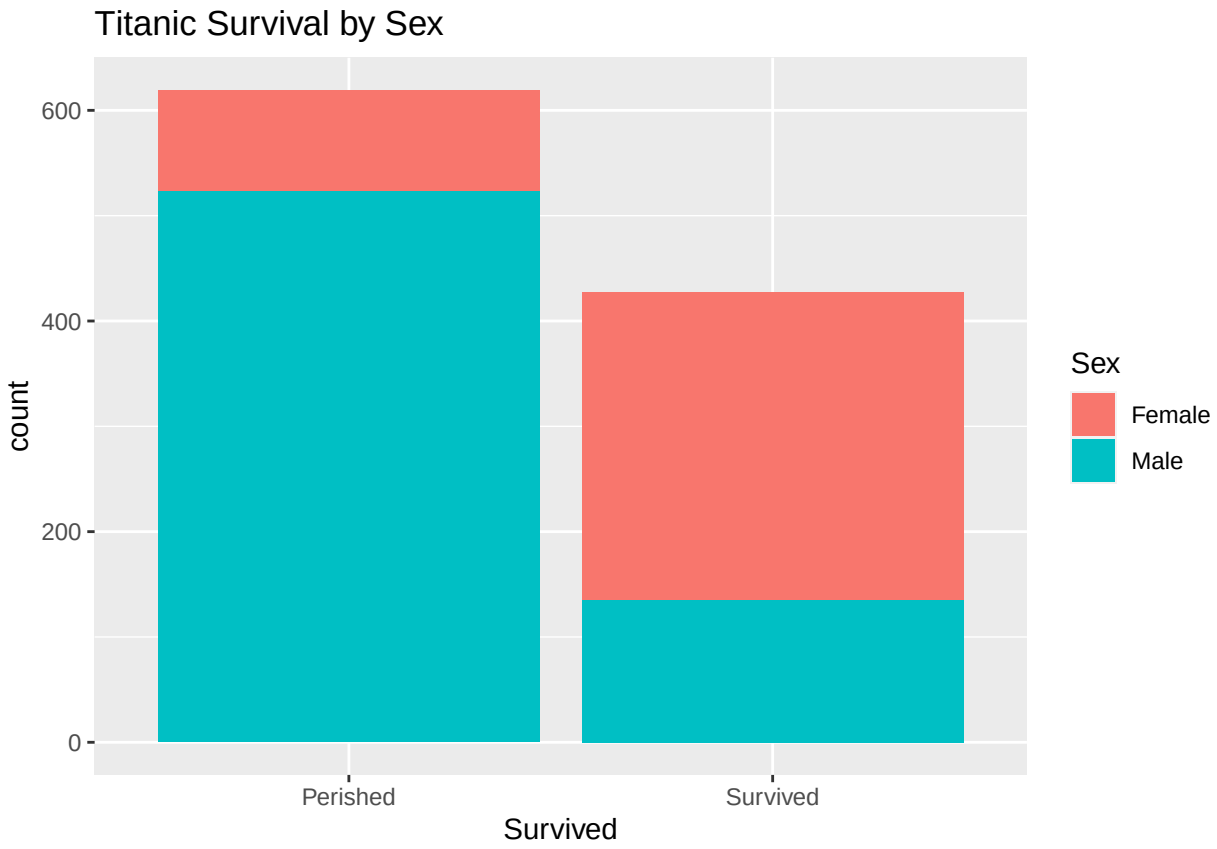
```
## Warning: package 'dplyr' was built under R version 4.0.4
```

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

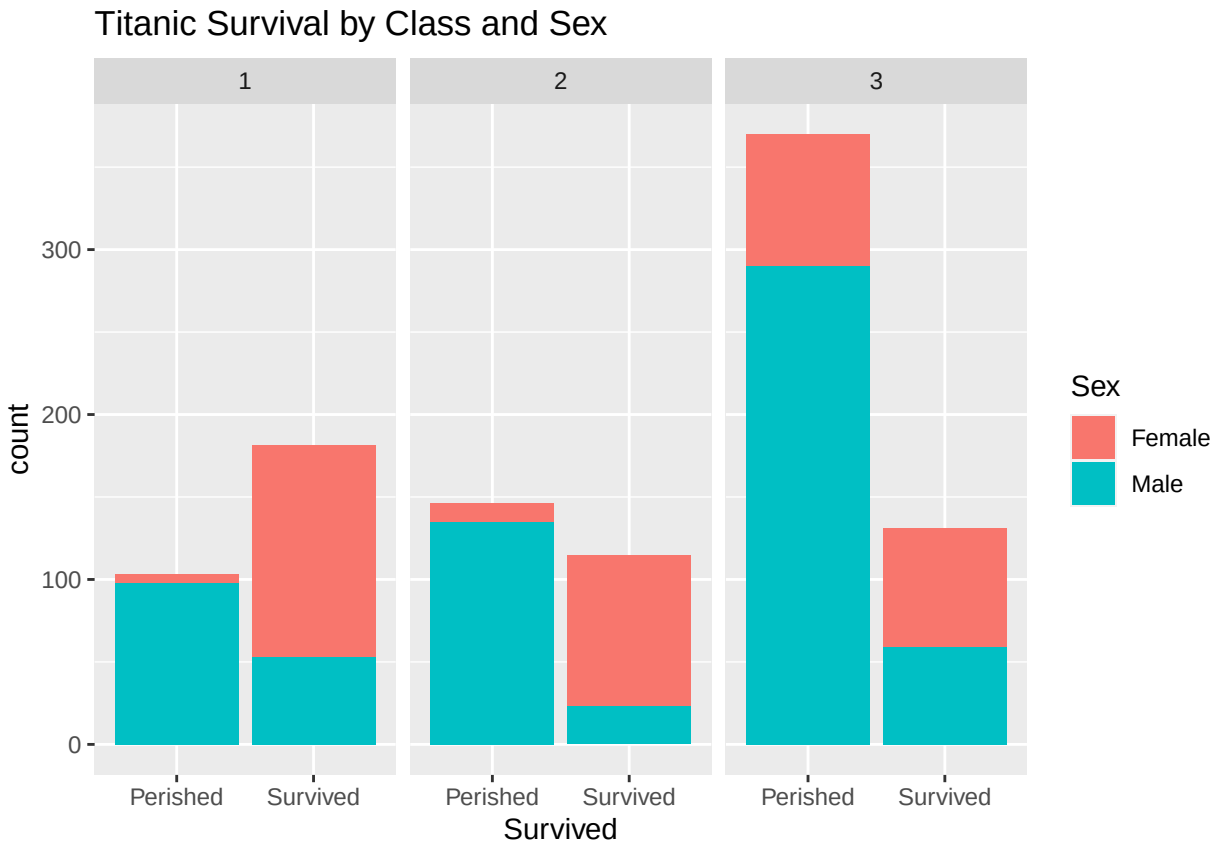
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

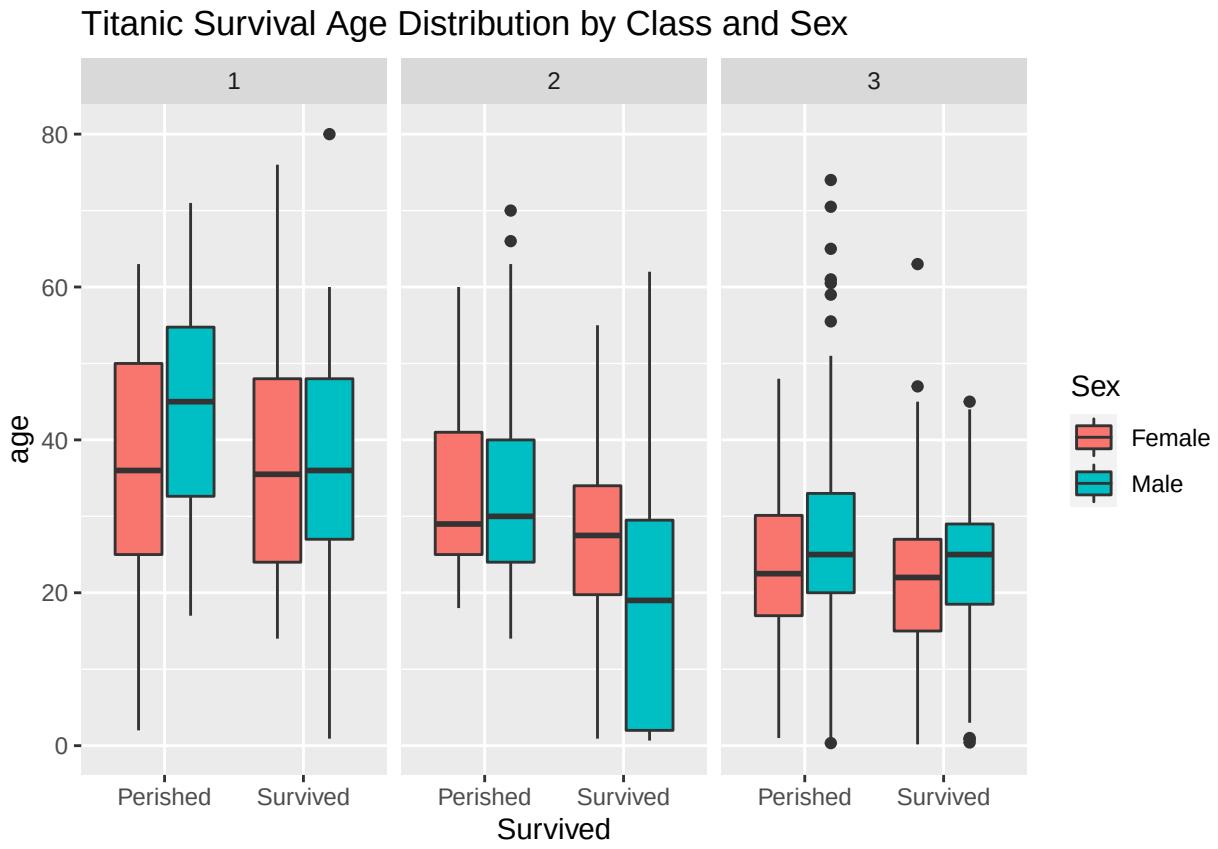
```
require(ggplot2)

#DATA EXPLORATION: GRAPHS 1-4, FUNCTION 4
#graph exploration 1
ggplot(data = ttnc, mapping = aes(x = as.factor(survived), fill = as.factor(sex))) +
  geom_bar() +
  labs(c("0", "1"), title = "Titanic Survival by Sex") + scale_x_discrete(name = "Survived", labels=c("
  scale_fill_discrete(name = "Sex", labels = c("Female", "Male"))
```



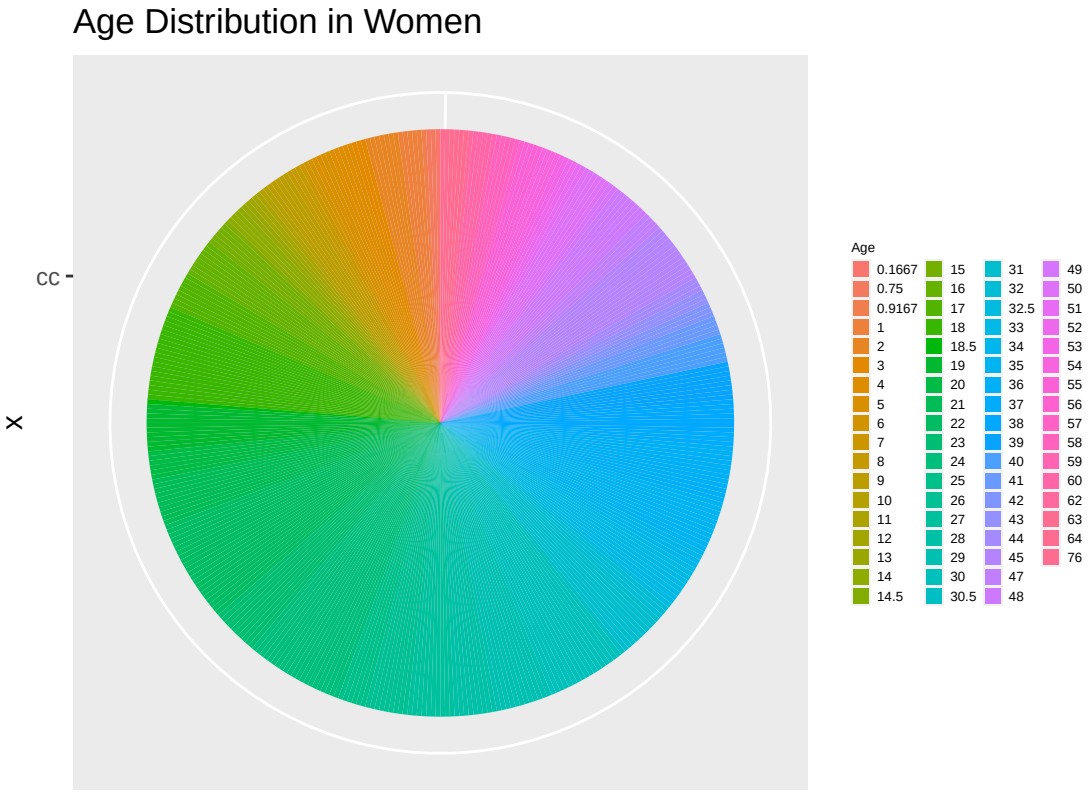

```
#graph exploration 2
ggplot(data = ttnc, mapping = aes(x = as.factor(survived), fill = as.factor(sex))) +
  geom_bar() +
  facet_wrap(~ pclass) +
  labs(c("0", "1"), title = "Titanic Survival by Class and Sex") + scale_x_discrete(name = "Survived",
  scale_fill_discrete(name = "Sex", labels = c("Female", "Male"))
```




```
#graph exploration 3
ggplot(data = ttnc, mapping = aes(as.factor(survived) , age, fill = as.factor(sex))) +
  geom_boxplot() +
  facet_wrap(~ pclass) +
  labs(c("0", "1"), title = "Titanic Survival Age Distribution by Class and Sex") + scale_x_discrete(na
  scale_fill_discrete(name = "Sex", labels = c("Female", "Male"))
```




```
#graph exploration 4
women.sub <- subset(ttnc, sex == 0)
ggplot(women.sub, aes(x="cc", y=as.factor(sex), fill=as.factor(age))) +
  geom_bar(stat="identity", width=1) +
  labs(title = "Age Distribution in Women", lab = "Women") +
  guides(fill=guide_legend(title="Age")) +
  coord_polar("y", start=0) + theme(
    legend.title = element_text(size = 5),
    legend.text = element_text(size = 5),
    legend.key.size = unit(.5, "line")
  )
```




```
#data exploration 4
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace, family = "binomial")
##
## A-priori probabilities:
## Y
##   0   1
## 0.6 0.4
##
## Conditional probabilities:
##    pclass
## Y          1         2         3
##   0 0.1685185 0.2203704 0.6111111
##   1 0.4166667 0.2638889 0.3194444
##
##    sex
## Y          0         1
##   0 0.1592593 0.8407407
##   1 0.6944444 0.3055556
##
##    age
```

```
## Y       [,1]     [,2]
##   0 30.41682 14.21185
##   1 28.92060 15.09074
```

```
##data exploration 5 + 6
head(ttnc)
```

```
##     X pclass survived sex age
## 1 738      3        0   1  19
## 2 868      3        1   0  22
## 3 971      3        1   1  20
## 4 938      3        0   0   1
## 5 456      2        0   1  63
## 6 139      1        0   1  38
```

```
tail(ttnc)
```

```
##         X pclass survived sex age
## 1041  789      3        0   1  45
## 1042  407      2        0   1  40
## 1043 1131      3        0   0  18
## 1044  953      3        0   1  22
## 1045  432      2        0   1  28
## 1046  756      3        0   1  17
```

```
dim(ttnc)
```

```
## [1] 1046    5
```