Assignment\_3

YASHWANTH REDDY SADALA

10-15-2023

knitr::opts\_chunk$set(echo = TRUE)

#Summary 1. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why? Ans: According to the data in this dataset, there is a 50.88% chance that an injury occurred if an accident has just been reported and no other information is available. This is because data indicates that earlier out of 42,183 cases, 21,462 cases had reported “injury=yes.”

1. Select the first 24 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER\_R and TRAF\_CON\_R. Create a pivot table that examines INJURY as a function of the two predictors for these 12 records. Use all three variables in the pivot table as rows/columns? 2.1:- Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors? Ans: For each of the six potential combinations of the predictors, the precise Bayes conditional probabilities of an injury (INJURY = Yes) are:

Predictor combination Probability WEATHER\_R = 1 and TRAF\_CON\_R = 1 0.6666667 WEATHER\_R = 2 and TRAF\_CON\_R = 0 0.1818182 WEATHER\_R = 1 and TRAF\_CON\_R = 1 0.0000000 WEATHER\_R = 2 and TRAF\_CON\_R = 1 0.0000000 WEATHER\_R = 1 and TRAF\_CON\_R = 2 0.0000000 WEATHER\_R = 2 and TRAF\_CON\_R = 2 1.0000000

2.2:-Classify the 24 accidents using these probabilities and a cutoff of 0.5? Ans: The 24 accidents are quantitatively classified using their probability and a cutoff of 0.5:

[0.6666667 0.1818182 0.0000000 0.0000000 0.6666667 0.1818182 0.1818182 0.6666667 0.1818182 0.1818182 0.1818182 0.0000000 0.6666667 0.6666667 0.6666667 0.6666667 0.1818182 0.1818182 0.1818182 0.1818182 0.6666667 0.6666667 1.0000000 0.1818182]

qualitatively are:

[“yes” “no” “no” “no” “yes” “no” “no” “yes” “no” “no” “no” “no” “yes” “yes” “yes” “yes” “no” “no” “no” “no” “yes” “yes” “yes” “no”]

2.3:-Compute manually the naive Bayes conditional probability of an injury given WEATHER\_R = 1 and TRAF\_CON\_R = 1? Ans: The probability or output was “0” when the naive Bayes conditional probability of an injury was manually calculated using WEATHER\_R = 1 and TRAF\_CON\_R = 1.

2.4:-Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent? Ans: Now that the naive Bayes classifier has been applied to the 24 records and two predictors, it has been discovered that the resultant classifications and rankings do not match those of the exact Bayes computation. This was discovered after checking the model output to acquire probabilities and classifications for all 24 records.

1. Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%)? 3.1:- Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix? Ans: Using the naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response) The Following Confusion Matrix and Statistics are obtained. The accuracy comes out to be 53.7%.

Confusion Matrix and Statistics

Reference

Prediction no yes no 3444 4866 yes 2947 5617

Accuracy : 0.537

3.2:- What is the overall error of the validation set? Ans: The overall error of the validation set is “46.3”.

# Problem Statement

The file accidentsFull.csv contains information on 42,183 actual automobile accidents in 2001 in the United States that involved one of three levels of injury: NO INJURY, INJURY, or FATALITY. For each accident, additional information is recorded, such as day of week, weather conditions, and road type. A firm might be interested in developing a system for quickly classifying the severity of an accident based on initial reports and associated data in the system (some of which rely on GPS-assisted reporting).

Our goal here is to predict whether an accident just reported will involve an injury (MAX\_SEV\_IR = 1 or 2) or will not (MAX\_SEV\_IR = 0). For this purpose, create a dummy variable called INJURY that takes the value “yes” if MAX\_SEV\_IR = 1 or 2, and otherwise “no.”

1. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?
2. Select the first 24 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER\_R and TRAF\_CON\_R. Create a pivot table that examines INJURY as a function of the two predictors for these 12 records. Use all three variables in the pivot table as rows/columns. 2.1:- Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors. 2.2:-Classify the 24 accidents using these probabilities and a cutoff of 0.5. 2.3:-Compute manually the naive Bayes conditional probability of an injury given WEATHER\_R = 1 and TRAF\_CON\_R = 1. 2.4:-Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?
3. Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%). 3.1:- Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix. 3.2:- What is the overall error of the validation set?

## Data Importing and Cleaning

Install and load the required package

library(e1071)  
library(caret)

## Loading required package: ggplot2

## Loading required package: lattice

library(ggplot2)

Q1. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?

accidents <- read.csv("C:\\Users\\yashw\\FML\\accidentsFull.csv")  
accidents$INJURY = ifelse(accidents$MAX\_SEV\_IR>0,"yes","no")  
injury\_table <- table(accidents$INJURY)  
injury\_table

##   
## no yes   
## 20721 21462

head(accidents)

## HOUR\_I\_R ALCHL\_I ALIGN\_I STRATUM\_R WRK\_ZONE WKDY\_I\_R INT\_HWY LGTCON\_I\_R  
## 1 0 2 2 1 0 1 0 3  
## 2 1 2 1 0 0 1 1 3  
## 3 1 2 1 0 0 1 0 3  
## 4 1 2 1 1 0 0 0 3  
## 5 1 1 1 0 0 1 0 3  
## 6 1 2 1 1 0 1 0 3  
## MANCOL\_I\_R PED\_ACC\_R RELJCT\_I\_R REL\_RWY\_R PROFIL\_I\_R SPD\_LIM SUR\_COND  
## 1 0 0 1 0 1 40 4  
## 2 2 0 1 1 1 70 4  
## 3 2 0 1 1 1 35 4  
## 4 2 0 1 1 1 35 4  
## 5 2 0 0 1 1 25 4  
## 6 0 0 1 0 1 70 4  
## TRAF\_CON\_R TRAF\_WAY VEH\_INVL WEATHER\_R INJURY\_CRASH NO\_INJ\_I PRPTYDMG\_CRASH  
## 1 0 3 1 1 1 1 0  
## 2 0 3 2 2 0 0 1  
## 3 1 2 2 2 0 0 1  
## 4 1 2 2 1 0 0 1  
## 5 0 2 3 1 0 0 1  
## 6 0 2 1 2 1 1 0  
## FATALITIES MAX\_SEV\_IR INJURY  
## 1 0 1 yes  
## 2 0 0 no  
## 3 0 0 no  
## 4 0 0 no  
## 5 0 0 no  
## 6 0 1 yes

probability\_injury <- (injury\_table["yes"] / sum(injury\_table))\*100  
probability\_injury

## yes   
## 50.87832

#converting factors from variables  
for (i in c(1:dim(accidents)[2])){  
 accidents[,i] <- as.factor(accidents[,i])  
}  
head(accidents,n=24)

## HOUR\_I\_R ALCHL\_I ALIGN\_I STRATUM\_R WRK\_ZONE WKDY\_I\_R INT\_HWY LGTCON\_I\_R  
## 1 0 2 2 1 0 1 0 3  
## 2 1 2 1 0 0 1 1 3  
## 3 1 2 1 0 0 1 0 3  
## 4 1 2 1 1 0 0 0 3  
## 5 1 1 1 0 0 1 0 3  
## 6 1 2 1 1 0 1 0 3  
## 7 1 2 1 0 0 1 1 3  
## 8 1 2 1 1 0 1 0 3  
## 9 1 2 1 1 0 1 0 3  
## 10 0 2 1 0 0 0 0 3  
## 11 1 2 1 0 0 1 0 3  
## 12 1 2 1 1 0 1 0 3  
## 13 1 2 1 1 0 1 0 3  
## 14 1 2 2 0 0 1 0 3  
## 15 1 2 2 1 0 1 0 3  
## 16 1 2 2 1 0 1 0 3  
## 17 1 2 1 1 0 1 0 3  
## 18 1 2 1 1 0 0 0 3  
## 19 1 2 1 1 0 1 0 3  
## 20 1 2 1 0 0 1 0 3  
## 21 1 2 1 1 0 1 0 3  
## 22 1 2 2 0 0 1 0 3  
## 23 1 2 1 0 0 1 0 3  
## 24 1 2 1 1 0 1 9 3  
## MANCOL\_I\_R PED\_ACC\_R RELJCT\_I\_R REL\_RWY\_R PROFIL\_I\_R SPD\_LIM SUR\_COND  
## 1 0 0 1 0 1 40 4  
## 2 2 0 1 1 1 70 4  
## 3 2 0 1 1 1 35 4  
## 4 2 0 1 1 1 35 4  
## 5 2 0 0 1 1 25 4  
## 6 0 0 1 0 1 70 4  
## 7 0 0 0 0 1 70 4  
## 8 0 0 0 0 1 35 4  
## 9 0 0 1 0 1 30 4  
## 10 0 0 1 0 1 25 4  
## 11 0 0 0 0 1 55 4  
## 12 2 0 0 1 1 40 4  
## 13 1 0 0 1 1 40 4  
## 14 0 0 0 0 1 25 4  
## 15 0 0 0 0 1 35 4  
## 16 0 0 0 0 1 45 4  
## 17 0 0 0 0 1 20 4  
## 18 0 0 0 0 1 50 4  
## 19 0 0 0 0 1 55 4  
## 20 0 0 1 1 1 55 4  
## 21 0 0 1 0 0 45 4  
## 22 0 0 1 0 0 65 4  
## 23 0 0 0 0 0 65 4  
## 24 2 0 1 1 0 55 4  
## TRAF\_CON\_R TRAF\_WAY VEH\_INVL WEATHER\_R INJURY\_CRASH NO\_INJ\_I PRPTYDMG\_CRASH  
## 1 0 3 1 1 1 1 0  
## 2 0 3 2 2 0 0 1  
## 3 1 2 2 2 0 0 1  
## 4 1 2 2 1 0 0 1  
## 5 0 2 3 1 0 0 1  
## 6 0 2 1 2 1 1 0  
## 7 0 2 1 2 0 0 1  
## 8 0 1 1 1 1 1 0  
## 9 0 1 1 2 0 0 1  
## 10 0 1 1 2 0 0 1  
## 11 0 1 1 2 0 0 1  
## 12 2 1 2 1 0 0 1  
## 13 0 1 4 1 1 2 0  
## 14 0 1 1 1 0 0 1  
## 15 0 1 1 1 1 1 0  
## 16 0 1 1 1 1 1 0  
## 17 0 1 1 2 0 0 1  
## 18 0 1 1 2 0 0 1  
## 19 0 1 1 2 0 0 1  
## 20 0 1 1 2 0 0 1  
## 21 0 3 1 1 1 1 0  
## 22 0 3 1 1 0 0 1  
## 23 2 2 1 2 1 2 0  
## 24 0 2 2 2 1 1 0  
## FATALITIES MAX\_SEV\_IR INJURY  
## 1 0 1 yes  
## 2 0 0 no  
## 3 0 0 no  
## 4 0 0 no  
## 5 0 0 no  
## 6 0 1 yes  
## 7 0 0 no  
## 8 0 1 yes  
## 9 0 0 no  
## 10 0 0 no  
## 11 0 0 no  
## 12 0 0 no  
## 13 0 1 yes  
## 14 0 0 no  
## 15 0 1 yes  
## 16 0 1 yes  
## 17 0 0 no  
## 18 0 0 no  
## 19 0 0 no  
## 20 0 0 no  
## 21 0 1 yes  
## 22 0 0 no  
## 23 0 1 yes  
## 24 0 1 yes

Q2. Select the first 24 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER\_R and TRAF\_CON\_R. Create a pivot table that examines INJURY as a function of the two predictors for these 12 records. Use all three variables in the pivot table as rows/columns. 2.1:- Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors. 2.2:-Classify the 24 accidents using these probabilities and a cutoff of 0.5. 2.3:=Compute manually the naive Bayes conditional probability of an injury given WEATHER\_R = 1 and TRAF\_CON\_R = 1. 2.4:-Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

accidents24 <- accidents[1:24,c("INJURY","WEATHER\_R","TRAF\_CON\_R")]  
head(accidents24)

## INJURY WEATHER\_R TRAF\_CON\_R  
## 1 yes 1 0  
## 2 no 2 0  
## 3 no 2 1  
## 4 no 1 1  
## 5 no 1 0  
## 6 yes 2 0

Pt1 <- ftable(accidents24)  
Pt2 <- ftable(accidents24[,-1]) # print table only for conditions  
Pt1

## TRAF\_CON\_R 0 1 2  
## INJURY WEATHER\_R   
## no 1 3 1 1  
## 2 9 1 0  
## yes 1 6 0 0  
## 2 2 0 1

Pt2

## TRAF\_CON\_R 0 1 2  
## WEATHER\_R   
## 1 9 1 1  
## 2 11 1 1

2.1:- Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.

#Injury = yes  
p1 = Pt1[3,1] / Pt2[1,1] # Injury, Weather=1 and Traf=0  
p2 = Pt1[4,1] / Pt2[2,1] # Injury, Weather=2, Traf=0  
p3 = Pt1[3,2] / Pt2[1,2] # Injury, W=1, T=1  
p4 = Pt1[4,2] / Pt2[2,2] # I, W=2,T=1  
p5 = Pt1[3,3] / Pt2[1,3] # I, W=1,T=2  
p6 = Pt1[4,3]/ Pt2[2,3] #I,W=2,T=2  
  
# Injury = no  
n1 = Pt1[1,1] / Pt2[1,1] # Weather=1 and Traf=0  
n2 = Pt1[2,1] / Pt2[2,1] # Weather=2, Traf=0  
n3 = Pt1[1,2] / Pt2[1,2] # W=1, T=1  
n4 = Pt1[2,2] / Pt2[2,2] # W=2,T=1  
n5 = Pt1[1,3] / Pt2[1,3] # W=1,T=2  
n6 = Pt1[2,3] / Pt2[2,3] # W=2,T=2  
print(c(p1,p2,p3,p4,p5,p6))

## [1] 0.6666667 0.1818182 0.0000000 0.0000000 0.0000000 1.0000000

print(c(n1,n2,n3,n4,n5,n6))

## [1] 0.3333333 0.8181818 1.0000000 1.0000000 1.0000000 0.0000000

2.2:-Classify the 24 accidents using these probabilities and a cutoff of 0.5.

prob.inj <- rep(0,24)  
  
for (i in 1:24) {  
 print(c(accidents24$WEATHER\_R[i],accidents24$TRAF\_CON\_R[i]))  
 if (accidents24$WEATHER\_R[i] == "1") {  
 if (accidents24$TRAF\_CON\_R[i]=="0"){  
 prob.inj[i] = p1  
 }  
 else if (accidents24$TRAF\_CON\_R[i]=="1") {  
 prob.inj[i] = p3  
 }  
 else if (accidents24$TRAF\_CON\_R[i]=="2") {  
 prob.inj[i] = p5  
 }  
 }  
 else {  
 if (accidents24$TRAF\_CON\_R[i]=="0"){  
 prob.inj[i] = p2  
 }  
 else if (accidents24$TRAF\_CON\_R[i]=="1") {  
 prob.inj[i] = p4  
 }  
 else if (accidents24$TRAF\_CON\_R[i]=="2") {  
 prob.inj[i] = p6  
 }  
 }  
 }

## [1] 1 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 2 1  
## Levels: 1 2 0  
## [1] 1 1  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 1 2  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 1 0  
## Levels: 1 2 0  
## [1] 2 2  
## Levels: 1 2 0  
## [1] 2 0  
## Levels: 1 2 0

accidents24$prob.inj <- prob.inj  
accidents24$prob.inj

## [1] 0.6666667 0.1818182 0.0000000 0.0000000 0.6666667 0.1818182 0.1818182  
## [8] 0.6666667 0.1818182 0.1818182 0.1818182 0.0000000 0.6666667 0.6666667  
## [15] 0.6666667 0.6666667 0.1818182 0.1818182 0.1818182 0.1818182 0.6666667  
## [22] 0.6666667 1.0000000 0.1818182

accidents24$pred.prob <- ifelse(accidents24$prob.inj>0.5, "yes", "no")  
accidents24$pred.prob

## [1] "yes" "no" "no" "no" "yes" "no" "no" "yes" "no" "no" "no" "no"   
## [13] "yes" "yes" "yes" "yes" "no" "no" "no" "no" "yes" "yes" "yes" "no"

2.3Compute manually the naive Bayes conditional probability of an injury given WEATHER\_R = 1 and TRAF\_CON\_R = 1. Answer:- Probability(Injury=Yes/WEATHER\_R=1,TRAF\_CON\_R=1)

= [ Probability(W=1/Injury=Yes) \* Probability(TRAF\_CON\_R=1/Injury=Yes) \* Probability(Injury=Yes) ] / [ Probability(W=1/Injury=Yes) \* Probability(TRAF\_CON\_R=1/Injury=Yes) \* Probability(Injury=Yes) + Probability(WEATHER\_R=1/Injury=No) \* Probability(TRAF\_CON\_R=1/Injury=No) \* Probability(Injury=No) ]

= [ 6/9 \* 0/9 \* 9/24 ] / [ 6/9 \* 0/9 \* 9/24 + 5/15 \* 2/15 \* 15/24 ] = The result will be “0” since the numerator is equal to zero.

2.4:- Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

nb <- naiveBayes(INJURY ~ TRAF\_CON\_R + WEATHER\_R,   
 data = accidents24)  
  
nbt <- predict(nb, newdata = accidents24,type = "raw")  
accidents24$nbpred.prob <- nbt[,2] # Transfer the "Yes" nb prediction  
accidents24$nbpred.prob

## [1] 0.571428571 0.250000000 0.002244949 0.008919722 0.571428571 0.250000000  
## [7] 0.250000000 0.571428571 0.250000000 0.250000000 0.250000000 0.666666667  
## [13] 0.571428571 0.571428571 0.571428571 0.571428571 0.250000000 0.250000000  
## [19] 0.250000000 0.250000000 0.571428571 0.571428571 0.333333333 0.250000000

Let us use Caret

library(klaR)

## Loading required package: MASS

#Loading the klaR package for Naive Bayes  
  
# Creating a variable named formula that includes all variables of interest  
formula <- INJURY ~ TRAF\_CON\_R + WEATHER\_R  
# Training the Naive Bayes model with Laplace  
  
accidents24$INJURY <- as.factor(accidents24$INJURY)  
nb2 <- NaiveBayes(formula,data = accidents24, laplace = 1)  
  
# Making predictions with the model  
 predict(nb2, newdata = accidents24[, c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")])

## $class  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20   
## yes no no no yes no no yes no no no yes yes yes yes yes no no no no   
## 21 22 23 24   
## yes yes no no   
## Levels: no yes  
##   
## $posterior  
## no yes  
## 1 0.4285714 0.571428571  
## 2 0.7500000 0.250000000  
## 3 0.9977551 0.002244949  
## 4 0.9910803 0.008919722  
## 5 0.4285714 0.571428571  
## 6 0.7500000 0.250000000  
## 7 0.7500000 0.250000000  
## 8 0.4285714 0.571428571  
## 9 0.7500000 0.250000000  
## 10 0.7500000 0.250000000  
## 11 0.7500000 0.250000000  
## 12 0.3333333 0.666666667  
## 13 0.4285714 0.571428571  
## 14 0.4285714 0.571428571  
## 15 0.4285714 0.571428571  
## 16 0.4285714 0.571428571  
## 17 0.7500000 0.250000000  
## 18 0.7500000 0.250000000  
## 19 0.7500000 0.250000000  
## 20 0.7500000 0.250000000  
## 21 0.4285714 0.571428571  
## 22 0.4285714 0.571428571  
## 23 0.6666667 0.333333333  
## 24 0.7500000 0.250000000

predict(nb2, newdata = accidents24[, c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")], type = "raw")

## $class  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20   
## yes no no no yes no no yes no no no yes yes yes yes yes no no no no   
## 21 22 23 24   
## yes yes no no   
## Levels: no yes  
##   
## $posterior  
## no yes  
## 1 0.4285714 0.571428571  
## 2 0.7500000 0.250000000  
## 3 0.9977551 0.002244949  
## 4 0.9910803 0.008919722  
## 5 0.4285714 0.571428571  
## 6 0.7500000 0.250000000  
## 7 0.7500000 0.250000000  
## 8 0.4285714 0.571428571  
## 9 0.7500000 0.250000000  
## 10 0.7500000 0.250000000  
## 11 0.7500000 0.250000000  
## 12 0.3333333 0.666666667  
## 13 0.4285714 0.571428571  
## 14 0.4285714 0.571428571  
## 15 0.4285714 0.571428571  
## 16 0.4285714 0.571428571  
## 17 0.7500000 0.250000000  
## 18 0.7500000 0.250000000  
## 19 0.7500000 0.250000000  
## 20 0.7500000 0.250000000  
## 21 0.4285714 0.571428571  
## 22 0.4285714 0.571428571  
## 23 0.6666667 0.333333333  
## 24 0.7500000 0.250000000

#predictions  
#raw\_probabilities

# Comparing the naive Bayes model and exact Bayes classification  
classification\_match <- all(accidents24$nbpred.prob == accidents24$prob.inj)  
probability\_match <- all.equal(accidents24$nbpred.prob, accidents24$prob.inj)  
  
# Checking if classifications and rankings are equivalent  
if (classification\_match && is.na(probability\_match)) {  
 cat("The resulting classifications and rankings are equivalent.\n")  
} else {  
 cat("The resulting classifications and rankings are not equivalent.\n")  
}

## The resulting classifications and rankings are not equivalent.

Q3, Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%). 3.1, Run a naive Bayes classifier on the complete training set with the relevant predictors(and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix.

set.seed(123)  
train.index <- sample(c(1:dim(accidents)[1]), dim(accidents)[1]\*0.6)   
train.df <- accidents[train.index,]  
valid.df <- accidents[-train.index,]  
#defining a variable to be used here  
vars <- c("INJURY", "HOUR\_I\_R", "ALIGN\_I" ,"WRK\_ZONE", "WKDY\_I\_R",  
 "INT\_HWY", "LGTCON\_I\_R", "PROFIL\_I\_R", "SPD\_LIM", "SUR\_COND",  
 "TRAF\_CON\_R", "TRAF\_WAY", "WEATHER\_R")  
  
nbTotal <- naiveBayes(INJURY~.,data = train.df[,vars])  
nbTotal

##   
## Naive Bayes Classifier for Discrete Predictors  
##   
## Call:  
## naiveBayes.default(x = X, y = Y, laplace = laplace)  
##   
## A-priori probabilities:  
## Y  
## no yes   
## 0.4903789 0.5096211   
##   
## Conditional probabilities:  
## HOUR\_I\_R  
## Y 0 1  
## no 0.5690919 0.4309081  
## yes 0.5690029 0.4309971  
##   
## ALIGN\_I  
## Y 1 2  
## no 0.8726936 0.1273064  
## yes 0.8696697 0.1303303  
##   
## WRK\_ZONE  
## Y 0 1  
## no 0.97502216 0.02497784  
## yes 0.97883393 0.02116607  
##   
## WKDY\_I\_R  
## Y 0 1  
## no 0.2190798 0.7809202  
## yes 0.2384091 0.7615909  
##   
## INT\_HWY  
## Y 0 1 9  
## no 0.8491660624 0.1501087745 0.0007251632  
## yes 0.8617615134 0.1374631726 0.0007753140  
##   
## LGTCON\_I\_R  
## Y 1 2 3  
## no 0.6871324 0.1285150 0.1843526  
## yes 0.6957668 0.1131958 0.1910374  
##   
## PROFIL\_I\_R  
## Y 0 1  
## no 0.7555394 0.2444606  
## yes 0.7617460 0.2382540  
##   
## SPD\_LIM  
## Y 5 10 15 20 25  
## no 8.057368e-05 7.251632e-04 4.673274e-03 8.299090e-03 1.099831e-01  
## yes 7.753140e-05 3.876570e-04 4.419290e-03 4.729415e-03 9.094433e-02  
## SPD\_LIM  
## Y 30 35 40 45 50  
## no 8.726130e-02 1.892676e-01 9.411006e-02 1.560712e-01 4.101201e-02  
## yes 8.885098e-02 2.163901e-01 1.076911e-01 1.554505e-01 3.806792e-02  
## SPD\_LIM  
## Y 55 60 65 70 75  
## no 1.604222e-01 3.545242e-02 6.711788e-02 3.948111e-02 6.043026e-03  
## yes 1.532020e-01 4.209955e-02 6.179253e-02 2.876415e-02 7.132889e-03  
##   
## SUR\_COND  
## Y 1 2 3 4 9  
## no 0.778341794 0.173797438 0.015550721 0.028120216 0.004189832  
## yes 0.815475267 0.153434641 0.011397116 0.015118623 0.004574353  
##   
## TRAF\_CON\_R  
## Y 0 1 2  
## no 0.6581259 0.1907985 0.1510757  
## yes 0.6217243 0.2203442 0.1579315  
##   
## TRAF\_WAY  
## Y 1 2 3  
## no 0.57360406 0.37426477 0.05213117  
## yes 0.56419600 0.39471236 0.04109164  
##   
## WEATHER\_R  
## Y 1 2  
## no 0.8411893 0.1588107  
## yes 0.8717631 0.1282369

#generating the confusion matrix using the train.df, the prediction and the classes  
confusionMatrix(train.df$INJURY, predict(nbTotal, train.df[, vars]), positive = "yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 5214 7197  
## yes 4475 8423  
##   
## Accuracy : 0.5388   
## 95% CI : (0.5327, 0.545)  
## No Information Rate : 0.6172   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0735   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.5392   
## Specificity : 0.5381   
## Pos Pred Value : 0.6530   
## Neg Pred Value : 0.4201   
## Prevalence : 0.6172   
## Detection Rate : 0.3328   
## Detection Prevalence : 0.5096   
## Balanced Accuracy : 0.5387   
##   
## 'Positive' Class : yes   
##

3.2, What is the overall error of the validation set?

confusionMatrix(valid.df$INJURY, predict(nbTotal, valid.df[, vars]), positive = "yes")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction no yes  
## no 3444 4866  
## yes 2947 5617  
##   
## Accuracy : 0.537   
## 95% CI : (0.5294, 0.5445)  
## No Information Rate : 0.6213   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.0706   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.5358   
## Specificity : 0.5389   
## Pos Pred Value : 0.6559   
## Neg Pred Value : 0.4144   
## Prevalence : 0.6213   
## Detection Rate : 0.3329   
## Detection Prevalence : 0.5075   
## Balanced Accuracy : 0.5374   
##   
## 'Positive' Class : yes   
##

#Calculated overall error  
  
ver=1-0.537  
verp=ver\*100  
paste("Overall Error: ",verp)

## [1] "Overall Error: 46.3"

#CONCLUSION

The Naive Bayes classifier was used firstly to predict injury outcomes in a data set with 24 records then to the entire data set with using two predictors both times.

Using the exact Bayes classifier for the first 24 records, we discover that the most risky combination for drivers is WEATHER\_CON=2,TRAF\_CON=0 because the likelihood for injury is maximal at “1” in this case.

The model’s accuracy on the training set was 53.7%, and its validation error was 46.3%, showing a modest level of predictive ability. However, it makes the assumption that the predictor variables are independent, which may not always be the case in real-world situations and might result in errors.But for classification and ranking, we can utilize the Naive Bayes classifier. Although Naive Bayes is a straightforward and useful