#############################

## Problem 8.2 Automobile Accidents. The file Accidents.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."

#After you load the data

#You want to create and insert a dummy variable called "INJURY" in the data

#based on the value of MAX\_SEV\_R

#Note use AccidentsFull.csv

accidents.df <- read.csv("AccidentsFull.csv")

#create and insert a dummy variable called "INJURY" in the data

accidents.df$INJURY <- ifelse(accidents.df$MAX\_SEV\_IR>0, "yes", "no")

head(accidents.df)

## 8.2.a 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?

# you can get this information by generating a table based on injury and simply determining the proportion yes to the sum of yes and no

#remember the following gives you the table

inj.tbl <- table(accidents.df$INJURY)

#probability of injury in accident – remember you can index the table values

#such as: inj.tbl["yes"])

#so the probability of injury is almost 50.87

## 8.2.b Select the first 12 records in the dataset and look only at the

##response (INJURY) and the two predictors WEATHER\_R and TRAF\_CON\_R.

##i. 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.

##ii. Compute the exact Bayes conditional probabilities of an injury

##(INJURY = Yes) given the six possible combinations of the predictors.

##iii. Classify the 12 accidents using these probabilities and a cutoff of 0.5.

##iv. Compute manually the naive Bayes conditional probability of an injury

##given WEATHER\_R = 1 and TRAF\_CON\_R = 1.

##v. Run a naive Bayes classifier on the 12 records and two predictors using R.

##Check the model output to obtain probabilities and classifications for all 12

##records. Compare this to the exact Bayes classification. Are the resulting

##classifications equivalent? Is the ranking (= ordering) of observations

##equivalent?

#Note you will first want to

#convert variables to categorical type

#you can use the as.factor function for that

#you will want to go through each variable using a for statement

for (i in c(1:dim(accidents.df)[2])){

accidents.df[,i] <- as.factor(accidents.df[,i])

}

str(accidents.df)

head(accidents.df[, c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")], 12)

#i.

#select first 12 records using 1:12 as the first index reference

#here is the example of pivot table

#pivot table

ftable(accidents.df[1:12, c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")])

# TRAF\_CON\_R 0 1 2

#INJURY WEATHER\_R

# no 1 1 1 1

# 2 5 1 0

# yes 1 2 0 0

# 2 1 0 0

#ii.

#To find P(Injury=yes|WEATHER\_R = 1, TRAF\_CON\_R =0):

#Numerator = (proportion of combination {WEATHER\_R =1, TRAF\_CON\_R = 0}when Injury =

# yes) \* (proportion of injuries in all cases)

#Denominator = proportion of combination {WEATHER\_R =1, TRAF\_CON\_R = 0} in

#all cases

#you want to reference the resulting table for this based on the three variables

numerator1 <- 2/3 \* 3/12

denominator1 <- 3/12

#the probability is given by ration

prob110 <- numerator1/denominator1

prob110

#so P(Injury=yes|WEATHER\_R = 1, TRAF\_CON\_R =0) = 0.667

#To find P(Injury=yes|WEATHER\_R = 1, TRAF\_CON\_R =1):

#Numerator = (proportion of combination {WEATHER\_R =1, TRAF\_CON\_R =1}when Injury =

# yes) \* (proportion of injuries in all cases)

#Denominator = proportion of combination {WEATHER\_R =1, TRAF\_CON\_R =1} in all cases

#again from the table of values we get

PUT YOUR Code here based on above

#so P(Injury=yes|WEATHER\_R = 1, TRAF\_CON\_R =1) = 0

#To find P(Injury=yes| WEATHER\_R = 1, TRAF\_CON\_R =2):

#Numerator = (proportion of combination {WEATHER\_R =1, TRAF\_CON\_R = 2}when Injury =

# yes) \* (proportion of injuries in all cases)

#Denominator =proportion of combination {WEATHER\_R =1, TRAF\_CON\_R = 2} in

#all cases

PUT YOUR Code here based on above

#so P(Injury=yes| WEATHER\_R = 1, TRAF\_CON\_R =2) = 0

#To find P(Injury=yes| WEATHER\_R = 2, TRAF\_CON\_R =0):

#Numerator = (proportion of combination {WEATHER\_R = 2, TRAF\_CON\_R = 0}when Injury =

# yes) \* (proportion of injuries in all cases)

#Denominator = proportion of combination {WEATHER\_R = 2, TRAF\_CON\_R = 0} in

#all cases

numerator4 <- 1/3 \* 3/12

denominator4 <- 6/12

prob4 <- numerator4/denominator4

prob4

#so P(Injury=yes| WEATHER\_R = 2, TRAF\_CON\_R =0) = 0.167

#To find P(Injury=yes| WEATHER\_R = 2, TRAF\_CON\_R =1):

#Numerator = (proportion of combination {WEATHER\_R = 2, TRAF\_CON\_R =1} when

# Injury = yes) \* (proportion of injuries in all cases)

#Denominator = proportion of combination {WEATHER\_R = 2, TRAF\_CON\_R =1} in

#all cases

PUT YOUR Code here based on above

#P(Injury=yes| WEATHER\_R = 2, TRAF\_CON\_R =1) = 0

#To find P(Injury=yes| WEATHER\_R = 2, TRAF\_CON\_R =2):

PUT YOUR Code here based on above

#In the above 12 observations there is no observation with

#(Injury=yes, WEATHER\_R = 2, TRAF\_CON\_R =2). The conditional probability here

#is undefined, since the denominator is zero.

#iii.

accidents <- head(accidents.df[, c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")], 12)

#At this point we have the probability of injuries vector and we add it as a column to the accidents data frame

prob.inj <- c(0.667, 0.167, 0, 0, 0.667, 0.167, 0.167, 0.667, 0.167, 0.167,

0.167, 0)

accidents$prob.inj <- prob.inj

accidents

#add a column for injury prediction based on cutoff of 0.5

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

accidents

#iv. Here we use the table of probabilities

#To find P(Injury=yes| WEATHER\_R = 1, TRAF\_CON\_R =1):

# Probability of injury involved in accidents

# = (proportion of WEATHER\_R =1 when Injury = yes)

# \*(proportion of TRAF\_CON\_R =1 when Injury = yes)

# \*(propotion of Injury = yes in all cases)

#

prob <- 2/3 \* 0/3 \* 3/12

prob

#so P(Injury=yes| WEATHER\_R = 1, TRAF\_CON\_R =1) = 0

#v. Using the basic form of naiveBayes

nb <- naiveBayes(INJURY ~ TRAF\_CON\_R + WEATHER\_R,

data = accidents.df[1:12, ])

predict(nb, newdata = accidents.df[1:12, c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")],

type = "raw")

#Provide your answer here

### to test nb with caret we use the train and predict functins

library(caret)

nb2 <- train(INJURY ~ TRAF\_CON\_R + WEATHER\_R,

data = accidents.df[1:12, ], method = "nb")

predict(nb2, newdata = accidents.df[1:12,

c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")])

predict(nb2, newdata = accidents.df[1:12,

c("INJURY", "WEATHER\_R", "TRAF\_CON\_R")],

type = "raw")

## 8.2.c Let us now return to the entire dataset. Partition the data into

##training (60%) and validation (40%).

#partiton the data into training (60%) and validation (40%) sets

#remember we use a proportion

set.seed(1)

train.index <- sample(c(1:dim(accidents.df)[1]), dim(accidents.df)[1]\*0.6)

valid.index <- setdiff(c(1:dim(accidents.df)[1]), train.index)

train.df <- accidents.df[train.index, ]

valid.df <- accidents.df[valid.index, ]

##8.2.c.i. Assuming that no information or initial reports about the accident itself

##are available at the time of prediction (only location characteristics,

##weather conditions, etc.), which predictors can we include in the analysis?

#Use the predictors are non-specific to the accident. They either

#describe calendar time or road conditions:

#HOUR\_I\_R, ALIGN\_I, WRK\_ZONE, WKDY\_I\_R, INT\_HWY, LGTCON\_I\_R, PROFIL\_I\_R,

#SPD\_LIM, SUR\_CON, TRAF\_CON\_R, TRAF\_WAY and WEATHER\_R.

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")

##8.2.c.ii. 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.

#fit the naive bayes model based on the model above we

#change the formula that we use all the variables and the data is train.df referenced the selected variables

nb <- naiveBayes(INJURY ~ ., data = train.df[, vars])

#generate the confusion matrix using the train.df, the prediction and the classes

#> confusionMatrix(train.df$INJURY, predict(nb, train.df[, vars]), positive = "yes")

#Confusion Matrix and Statistics

#

# Reference

# Prediction no yes

# no 5097 7405

# yes 4230 8577

#

# Accuracy : 0.5403

# 95% CI : (0.5341, 0.5464)

# No Information Rate : 0.6315

# P-Value [Acc > NIR] : 1

#

# Kappa : 0.0776

#

# Mcnemar's Test P-Value : <2e-16

#

# Sensitivity : 0.5367

# Specificity : 0.5465

# Pos Pred Value : 0.6697

# Neg Pred Value : 0.4077

# Prevalence : 0.6315

# Detection Rate : 0.3389

# Detection Prevalence : 0.5060

# Balanced Accuracy : 0.5416

#

# 'Positive' Class : yes

##8.2.c.iii. What is the overall error for the validation set?

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

#this is determined from the Confusion Matrix and Statistics

#So the Overall error for the validation set is 46.68%.

##8.2.c.iv. What is the percent improvement relative to the naive rule (using the

##validation set)?

#Overall error using validation set 0.4668

#Naive rule's error 0.5087

#calculate the improvement

#The improvement is 0.0419, or 8.23% better than the naive rule error.

##8.2.c.v. Examine the conditional probabilities output. Why do we get a probability

##of zero for P(INJURY = No j SPD\_LIM = 5)?

#review the table below

options(digits = 2)

nb

# SPD\_LIM

#Y 5 10 15 20 25 30 35 40 45

# no 0.00008 0.00048 0.00440 0.00856 0.11214 0.08607 0.18965 0.09622 0.15534

# yes 0.00016 0.00031 0.00406 0.00390 0.09065 0.08605 0.21231 0.10689 0.15741

# SPD\_LIM

#Y 50 55 60 65 70 75

# no 0.04079 0.15901 0.03551 0.06455 0.04095 0.00624

# yes 0.03943 0.15492 0.04302 0.06215 0.03115 0.00757