Drug

Zhe Zhao(433707)

2022/5/26

library(class)  
library(dplyr)

## Warning: łĚĽ­°ü'dplyr'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

##   
## 载入程辑包：'dplyr'

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

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

library(readr)

## Warning: łĚĽ­°ü'readr'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

library(ggplot2)

## Warning: łĚĽ­°ü'ggplot2'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

library(caret)

## 载入需要的程辑包：lattice

library(verification)

## Warning: łĚĽ­°ü'verification'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

## 载入需要的程辑包：fields

## Warning: łĚĽ­°ü'fields'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

## 载入需要的程辑包：spam

## Warning: łĚĽ­°ü'spam'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

## Spam version 2.8-0 (2022-01-05) is loaded.  
## Type 'help( Spam)' or 'demo( spam)' for a short introduction   
## and overview of this package.  
## Help for individual functions is also obtained by adding the  
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.

##   
## 载入程辑包：'spam'

## The following objects are masked from 'package:base':  
##   
## backsolve, forwardsolve

## 载入需要的程辑包：viridis

## Warning: łĚĽ­°ü'viridis'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

## 载入需要的程辑包：viridisLite

##   
## Try help(fields) to get started.

## 载入需要的程辑包：boot

##   
## 载入程辑包：'boot'

## The following object is masked from 'package:lattice':  
##   
## melanoma

## 载入需要的程辑包：CircStats

## Warning: łĚĽ­°ü'CircStats'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

## 载入需要的程辑包：MASS

##   
## 载入程辑包：'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

## 载入需要的程辑包：dtw

## Warning: łĚĽ­°ü'dtw'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

## 载入需要的程辑包：proxy

##   
## 载入程辑包：'proxy'

## The following object is masked from 'package:spam':  
##   
## as.matrix

## The following objects are masked from 'package:stats':  
##   
## as.dist, dist

## The following object is masked from 'package:base':  
##   
## as.matrix

## Loaded dtw v1.22-3. See ?dtw for help, citation("dtw") for use in publication.

## Registered S3 method overwritten by 'verification':  
## method from  
## lines.roc pROC

library(corrplot)

## corrplot 0.92 loaded

library(gmodels)  
library(kernlab)

## Warning: łĚĽ­°ü'kernlab'ĘÇÓĂR°ć±ľ4.1.3 Ŕ´˝¨ÔěµÄ

##   
## 载入程辑包：'kernlab'

## The following object is masked from 'package:CircStats':  
##   
## rvm

## The following object is masked from 'package:ggplot2':  
##   
## alpha

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble 3.1.6 v stringr 1.4.0  
## v tidyr 1.1.4 v forcats 0.5.1  
## v purrr 0.3.4

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x kernlab::alpha() masks ggplot2::alpha()  
## x purrr::cross() masks kernlab::cross()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x purrr::lift() masks caret::lift()  
## x MASS::select() masks dplyr::select()

## Introduction

The main target of this project is using classification method of machine learning, to build a model explaining whether a particular person consumed cocaine in the last month based on the training sample and generate predictions for all observations from the test sample.

## Import data

Import data and check detailed information.

## Read data  
drug <- read.csv("C:/Users/zhaoz/Desktop/jiqixuexi/pro/drugs\_train.csv")  
  
## get a quick peek at the data  
head(drug)

## id age gender education  
## 1 train\_0001 45-54 male Masters degree  
## 2 train\_0002 25-34 male University degree  
## 3 train\_0003 18-24 female University degree  
## 4 train\_0004 25-34 female Masters degree  
## 5 train\_0005 18-24 male Some college or university, no certificate or degree  
## 6 train\_0006 18-24 female Left school at 18 years  
## country ethnicity personality\_neuroticism personality\_extraversion  
## 1 USA Mixed-Black/Asian 57.6 57.3  
## 2 USA Mixed-Black/Asian 47.8 67.0  
## 3 USA Mixed-Black/Asian 57.6 43.3  
## 4 USA Mixed-Black/Asian 71.8 31.2  
## 5 Australia Mixed-Black/Asian 56.1 62.3  
## 6 Australia Mixed-White/Black 47.8 20.7  
## personality\_openness personality\_agreeableness personality\_conscientiousness  
## 1 50.1 47.8 53.7  
## 2 45.7 47.8 56.0  
## 3 55.3 45.6 49.9  
## 4 43.6 56.3 31.8  
## 5 70.2 66.1 42.4  
## 6 57.8 41.2 33.6  
## personality\_impulsiveness personality\_sensation consumption\_alcohol  
## 1 42.8 22.4 used in last week  
## 2 33.8 30.8 used in last week  
## 3 63.0 62.0 used in last month  
## 4 63.0 71.1 used in last day  
## 5 50.4 62.0 used in last week  
## 6 56.5 62.0 used in last year  
## consumption\_amphetamines consumption\_caffeine consumption\_cannabis  
## 1 used over a decade ago used in last day used in last week  
## 2 never used used in last week never used  
## 3 never used used in last day used in last week  
## 4 never used used in last day used in last decade  
## 5 never used used in last month used in last month  
## 6 never used used in last day used in last month  
## consumption\_chocolate consumption\_mushrooms consumption\_nicotine  
## 1 used in last day never used used in last week  
## 2 used in last day never used never used  
## 3 used in last week used in last year used in last month  
## 4 used in last day never used used in last decade  
## 5 used in last day used in last year used in last month  
## 6 used in last month used in last year used in last decade  
## consumption\_cocaine\_last\_month  
## 1 No  
## 2 No  
## 3 No  
## 4 No  
## 5 No  
## 6 No

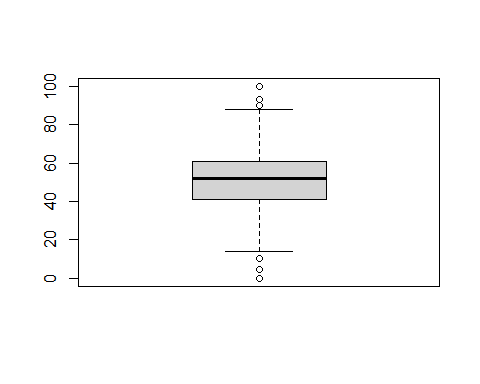
## get dataset detailed infomation  
str(drug)

## 'data.frame': 1500 obs. of 21 variables:  
## $ id : chr "train\_0001" "train\_0002" "train\_0003" "train\_0004" ...  
## $ age : chr "45-54" "25-34" "18-24" "25-34" ...  
## $ gender : chr "male" "male" "female" "female" ...  
## $ education : chr "Masters degree" "University degree" "University degree" "Masters degree" ...  
## $ country : chr "USA" "USA" "USA" "USA" ...  
## $ ethnicity : chr "Mixed-Black/Asian" "Mixed-Black/Asian" "Mixed-Black/Asian" "Mixed-Black/Asian" ...  
## $ personality\_neuroticism : num 57.6 47.8 57.6 71.8 56.1 47.8 75.2 53.4 60.8 30.1 ...  
## $ personality\_extraversion : num 57.3 67 43.3 31.2 62.3 20.7 69.6 54.9 72.2 93.7 ...  
## $ personality\_openness : num 50.1 45.7 55.3 43.6 70.2 57.8 62.5 27.8 62.5 64.7 ...  
## $ personality\_agreeableness : num 47.8 47.8 45.6 56.3 66.1 41.2 54.2 70.9 51.9 82.2 ...  
## $ personality\_conscientiousness : num 53.7 56 49.9 31.8 42.4 33.6 38.7 66.4 63.6 53.7 ...  
## $ personality\_impulsiveness : num 42.8 33.8 63 63 50.4 56.5 56.5 21.5 50.4 56.5 ...  
## $ personality\_sensation : num 22.4 30.8 62 71.1 62 62 22.4 38.8 100 100 ...  
## $ consumption\_alcohol : chr "used in last week" "used in last week" "used in last month" "used in last day" ...  
## $ consumption\_amphetamines : chr "used over a decade ago" "never used" "never used" "never used" ...  
## $ consumption\_caffeine : chr "used in last day" "used in last week" "used in last day" "used in last day" ...  
## $ consumption\_cannabis : chr "used in last week" "never used" "used in last week" "used in last decade" ...  
## $ consumption\_chocolate : chr "used in last day" "used in last day" "used in last week" "used in last day" ...  
## $ consumption\_mushrooms : chr "never used" "never used" "used in last year" "never used" ...  
## $ consumption\_nicotine : chr "used in last week" "never used" "used in last month" "used in last decade" ...  
## $ consumption\_cocaine\_last\_month: chr "No" "No" "No" "No" ...

## Data Visualization

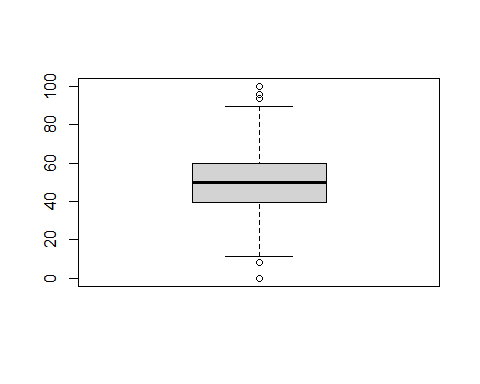
Plot a boxplot of variable personality\_neuroticism

boxplot(drug$personality\_neuroticism)



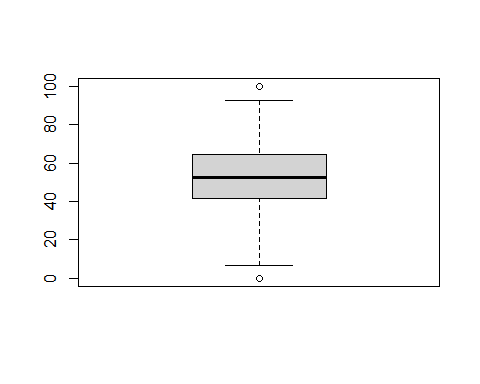
Plot a boxplot of variable personality\_extraversion

boxplot(drug$personality\_extraversion)



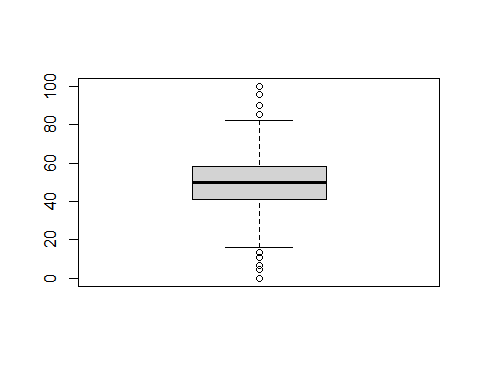
Plot a boxplot of variable personality\_openness

boxplot(drug$personality\_openness)



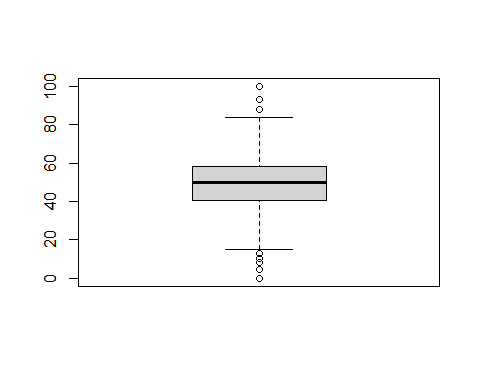
Plot a boxplot of variable personality\_agreeableness

boxplot(drug$personality\_agreeableness)



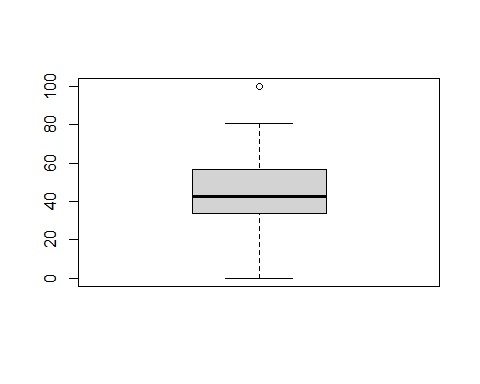
Plot a boxplot of variable personality\_conscientiousness

boxplot(drug$personality\_conscientiousness)



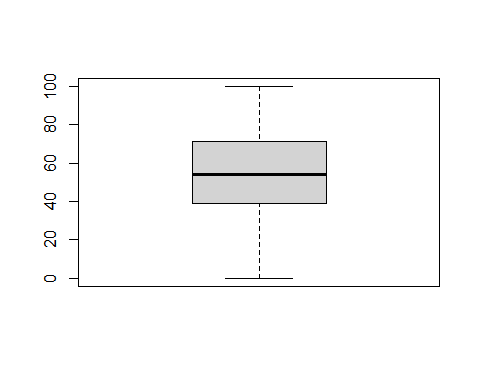
Plot a boxplot of variable personality\_impulsiveness

boxplot(drug$personality\_impulsiveness)



Plot a boxplot of variable personality\_sensation

boxplot(drug$personality\_sensation)



Divide the data into training and testing sample.

set.seed(123456)  
# divide the data into training and testing sample  
drug\_which\_training <- createDataPartition(drug$consumption\_cocaine\_last\_month,  
 p = 0.7,   
 list = FALSE)   
  
drug\_train <-drug[c(drug\_which\_training),]  
drug\_test <- drug[-c(drug\_which\_training),]  
## save data  
save(list = c("drug\_train",  
 "drug\_test"),  
 file = "C:/Users/zhaoz/Desktop/jiqixuexi/pro/drug\_train\_test.RData")

Check if there are any NA values.

any(is.na(drug\_train))

## [1] FALSE

## Data pre-processing

# convert all character variables to factors (apart from the 1st column - id)  
  
# create a vector of names  
drug\_categorical\_vars <-   
 sapply(drug\_train[, -1], is.character) %>%   
 which() %>%   
 names()  
  
drug\_categorical\_vars

## [1] "age" "gender"   
## [3] "education" "country"   
## [5] "ethnicity" "consumption\_alcohol"   
## [7] "consumption\_amphetamines" "consumption\_caffeine"   
## [9] "consumption\_cannabis" "consumption\_chocolate"   
## [11] "consumption\_mushrooms" "consumption\_nicotine"   
## [13] "consumption\_cocaine\_last\_month"

Apply a conversion in a loop and check all factor variables

# apply a conversion in a loop  
for (variable in drug\_categorical\_vars) {  
 drug\_train[[variable]] <- as.factor(drug\_train[[variable]])  
}  
## check all factor variables  
glimpse(drug\_train)

## Rows: 1,051  
## Columns: 21  
## $ id <chr> "train\_0002", "train\_0003", "train\_0004~  
## $ age <fct> 25-34, 18-24, 25-34, 18-24, 25-34, 25-3~  
## $ gender <fct> male, female, female, female, male, mal~  
## $ education <fct> "University degree", "University degree~  
## $ country <fct> USA, USA, USA, Australia, USA, USA, Aus~  
## $ ethnicity <fct> Mixed-Black/Asian, Mixed-Black/Asian, M~  
## $ personality\_neuroticism <dbl> 47.8, 57.6, 71.8, 47.8, 75.2, 53.4, 60.~  
## $ personality\_extraversion <dbl> 67.0, 43.3, 31.2, 20.7, 69.6, 54.9, 72.~  
## $ personality\_openness <dbl> 45.7, 55.3, 43.6, 57.8, 62.5, 27.8, 62.~  
## $ personality\_agreeableness <dbl> 47.8, 45.6, 56.3, 41.2, 54.2, 70.9, 51.~  
## $ personality\_conscientiousness <dbl> 56.0, 49.9, 31.8, 33.6, 38.7, 66.4, 63.~  
## $ personality\_impulsiveness <dbl> 33.8, 63.0, 63.0, 56.5, 56.5, 21.5, 50.~  
## $ personality\_sensation <dbl> 30.8, 62.0, 71.1, 62.0, 22.4, 38.8, 100~  
## $ consumption\_alcohol <fct> used in last week, used in last month, ~  
## $ consumption\_amphetamines <fct> never used, never used, never used, nev~  
## $ consumption\_caffeine <fct> used in last week, used in last day, us~  
## $ consumption\_cannabis <fct> never used, used in last week, used in ~  
## $ consumption\_chocolate <fct> used in last day, used in last week, us~  
## $ consumption\_mushrooms <fct> never used, used in last year, never us~  
## $ consumption\_nicotine <fct> never used, used in last month, used in~  
## $ consumption\_cocaine\_last\_month <fct> No, No, No, No, Yes, No, Yes, No, No, Y~

Verify transformation in data frame details, lets see if there are wrong type data in the training sample.

# verify transformation in data frame details  
str(drug\_train)

## 'data.frame': 1051 obs. of 21 variables:  
## $ id : chr "train\_0002" "train\_0003" "train\_0004" "train\_0006" ...  
## $ age : Factor w/ 6 levels "18-24","25-34",..: 2 1 2 1 2 2 1 3 1 1 ...  
## $ gender : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 1 1 1 1 ...  
## $ education : Factor w/ 9 levels "Doctorate degree",..: 9 9 6 4 1 1 9 4 3 8 ...  
## $ country : Factor w/ 7 levels "Australia","Canada",..: 7 7 7 1 7 7 1 7 5 1 ...  
## $ ethnicity : Factor w/ 7 levels "Asian","Black",..: 3 3 3 5 4 3 3 3 3 3 ...  
## $ personality\_neuroticism : num 47.8 57.6 71.8 47.8 75.2 53.4 60.8 42.8 49.2 63.7 ...  
## $ personality\_extraversion : num 67 43.3 31.2 20.7 69.6 54.9 72.2 25.1 50.1 33.3 ...  
## $ personality\_openness : num 45.7 55.3 43.6 57.8 62.5 27.8 62.5 21 62.5 62.5 ...  
## $ personality\_agreeableness : num 47.8 45.6 56.3 41.2 54.2 70.9 51.9 63.6 45.6 61 ...  
## $ personality\_conscientiousness : num 56 49.9 31.8 33.6 38.7 66.4 63.6 56 28.1 33.6 ...  
## $ personality\_impulsiveness : num 33.8 63 63 56.5 56.5 21.5 50.4 56.5 70.5 63 ...  
## $ personality\_sensation : num 30.8 62 71.1 62 22.4 38.8 100 54 82.6 38.8 ...  
## $ consumption\_alcohol : Factor w/ 7 levels "never used","used in last day",..: 5 4 2 6 2 6 2 2 5 2 ...  
## $ consumption\_amphetamines : Factor w/ 7 levels "never used","used in last day",..: 1 1 1 1 3 7 5 6 3 2 ...  
## $ consumption\_caffeine : Factor w/ 7 levels "never used","used in last day",..: 5 2 2 2 2 6 2 4 2 5 ...  
## $ consumption\_cannabis : Factor w/ 7 levels "never used","used in last day",..: 1 5 3 4 3 7 5 4 6 4 ...  
## $ consumption\_chocolate : Factor w/ 7 levels "never used","used in last day",..: 2 5 2 4 4 2 6 5 2 5 ...  
## $ consumption\_mushrooms : Factor w/ 7 levels "never used","used in last day",..: 1 6 1 6 1 1 6 1 6 4 ...  
## $ consumption\_nicotine : Factor w/ 7 levels "never used","used in last day",..: 1 4 3 3 2 7 2 2 2 4 ...  
## $ consumption\_cocaine\_last\_month: Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 1 2 ...

White a function to generate contingency table.

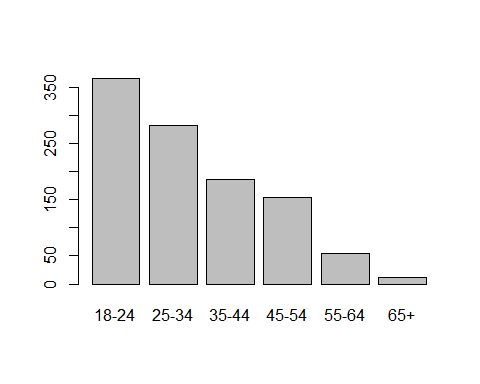
# function: generate contingency table  
get.contingency.table <- function(dep.var, indep.var, stat.tests=F){  
 if(stat.tests == F){  
 CrossTable(dep.var, indep.var, digits=1,  
 prop.r=F, prop.t=F, prop.chisq=F)  
 }else{  
 CrossTable(dep.var, indep.var, digits=1,  
 prop.r=F, prop.t=F, prop.chisq=F,  
 chisq=T, fisher=T)  
 }  
}

## Significant test

For numeric variables, I prefer to use t-test and for others, I use chi-square test and fisher test.

Draw a plot of variable age.

plot(drug\_train$age)

 Check levels and label them for analysis.

levels(drug\_train$age)

## [1] "18-24" "25-34" "35-44" "45-54" "55-64" "65+"

drug\_train$age<-factor(drug\_train$age,  
 level=c("18-24",  
 "25-34",  
 "35-44",  
 "45-54",  
 "55-64",  
 "65+"),  
 labels = c(1:6))

Make contingency table and do Chi-square test and fisher test of variable age.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$age)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 315 | 254 | 179 | 149 | 54 | 11 | 962 |   
## | 0.9 | 0.9 | 1.0 | 1.0 | 1.0 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 50 | 28 | 6 | 4 | 1 | 0 | 89 |   
## | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 365 | 282 | 185 | 153 | 55 | 11 | 1051 |   
## | 0.3 | 0.3 | 0.2 | 0.1 | 0.1 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$age, workspace = 2e7)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$age  
## p-value = 5.215e-06  
## alternative hypothesis: two.sided

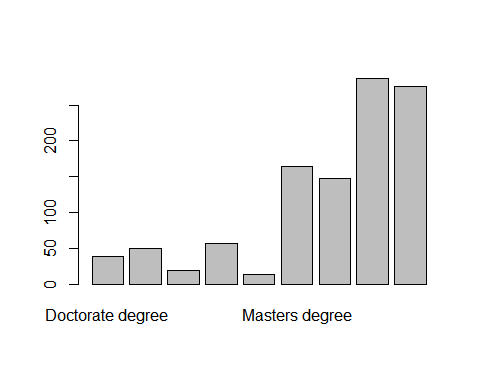
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$age)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$age): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$age  
## X-squared = 31.095, df = 5, p-value = 8.972e-06

Draw a plot of variable education.

plot(drug\_train$education)

 Check levels and label them for analysis.

levels(drug\_train$education)

## [1] "Doctorate degree"   
## [2] "Left school at 16 years"   
## [3] "Left school at 17 years"   
## [4] "Left school at 18 years"   
## [5] "Left school before 16 years"   
## [6] "Masters degree"   
## [7] "Professional certificate/ diploma"   
## [8] "Some college or university, no certificate or degree"  
## [9] "University degree"

drug\_train$education<-factor(drug\_train$education,  
 level=c("Left school before 16 years",   
 "Left school at 16 years",  
 "Left school at 17 years",   
 "Left school at 18 years",   
 "Some college or university, no certificate or degree",  
 "Professional certificate/ diploma",   
 "University degree",  
 "Masters degree",   
 "Doctorate degree"),  
 labels = c(1:9))

Make contingency table and do Chi-square test and fisher test of variable education.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$education)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 12 | 48 | 17 | 51 | 256 | 136 | 258 | 150 | 34 | 962 |   
## | 0.9 | 1.0 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 1 | 2 | 2 | 6 | 31 | 11 | 18 | 14 | 4 | 89 |   
## | 0.1 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 13 | 50 | 19 | 57 | 287 | 147 | 276 | 164 | 38 | 1051 |   
## | 0.0 | 0.0 | 0.0 | 0.1 | 0.3 | 0.1 | 0.3 | 0.2 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$education, workspace = 2e7)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$education  
## p-value = 0.6399  
## alternative hypothesis: two.sided

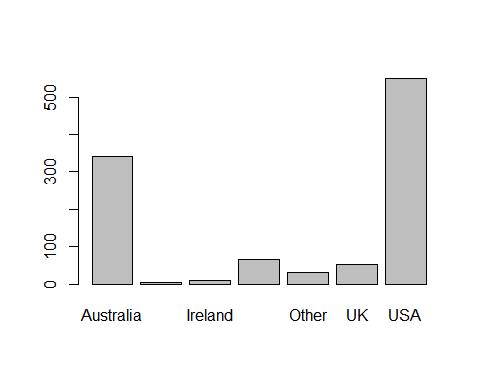
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$education)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$education): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$education  
## X-squared = 5.4708, df = 8, p-value = 0.7063

Draw a plot of variable country.

plot(drug\_train$country)

 Check levels and label them for analysis.

drug\_train$country<-factor(drug\_train$country,  
 level=c("Australia",  
 "Canada",   
 "Ireland",  
 "New Zealand",  
 "Other",  
 "UK",  
 "USA"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable country.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$country)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 300 | 4 | 8 | 58 | 28 | 45 | 519 | 962 |   
## | 0.9 | 1.0 | 0.9 | 0.9 | 0.9 | 0.8 | 0.9 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 40 | 0 | 1 | 8 | 2 | 8 | 30 | 89 |   
## | 0.1 | 0.0 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 340 | 4 | 9 | 66 | 30 | 53 | 549 | 1051 |   
## | 0.3 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 0.5 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$country)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$country  
## p-value = 0.008128  
## alternative hypothesis: two.sided

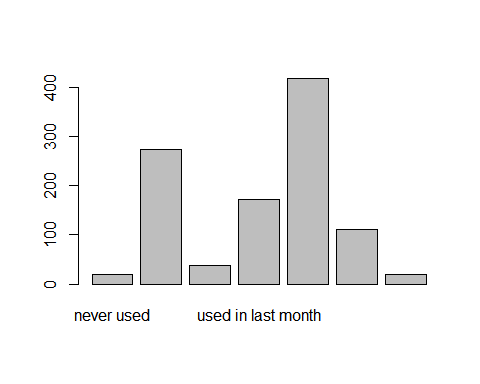
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$country)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$country): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$country  
## X-squared = 15.873, df = 6, p-value = 0.01445

Draw a plot of variable consumption\_alcohol.

plot(drug\_train$consumption\_alcohol)

 Label them for analysis.

drug\_train$consumption\_alcohol<-factor(drug\_train$consumption\_alcohol,  
 level=c("never used",  
 "used in last day",   
 "used in last week",  
 "used in last month",  
 "used in last year",  
 "used in last decade",  
 "used over a decade ago"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable consumption\_alcohol.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_alcohol)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 19 | 246 | 373 | 161 | 108 | 36 | 19 | 962 |   
## | 1.0 | 0.9 | 0.9 | 0.9 | 1.0 | 1.0 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 0 | 28 | 45 | 11 | 4 | 1 | 0 | 89 |   
## | 0.0 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 19 | 274 | 418 | 172 | 112 | 37 | 19 | 1051 |   
## | 0.0 | 0.3 | 0.4 | 0.2 | 0.1 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_alcohol, workspace = 2e7)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_alcohol  
## p-value = 0.0493  
## alternative hypothesis: two.sided

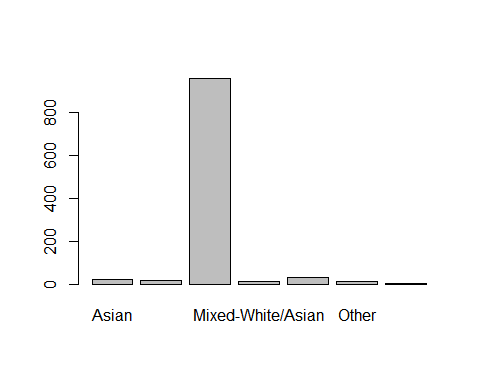
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_alcohol)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$consumption\_alcohol): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_alcohol  
## X-squared = 13.45, df = 6, p-value = 0.03641

Draw a plot of variable ethnicity.

plot(drug\_train$ethnicity)

 Label them for analysis.

drug\_train$ethnicity <- factor(drug\_train$ethnicity,  
 level=c("Asian",  
 "Black",  
 "Mixed-Black/Asian",  
 "Mixed-White/Asian",  
 "Mixed-White/Black",  
 "Other",  
 "White"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable ethnicity.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$ethnicity)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 20 | 15 | 881 | 6 | 27 | 10 | 3 | 962 |   
## | 0.9 | 1.0 | 0.9 | 0.6 | 0.8 | 0.9 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 2 | 0 | 77 | 4 | 5 | 1 | 0 | 89 |   
## | 0.1 | 0.0 | 0.1 | 0.4 | 0.2 | 0.1 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 22 | 15 | 958 | 10 | 32 | 11 | 3 | 1051 |   
## | 0.0 | 0.0 | 0.9 | 0.0 | 0.0 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$ethnicity, workspace = 2e7)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$ethnicity  
## p-value = 0.03333  
## alternative hypothesis: two.sided

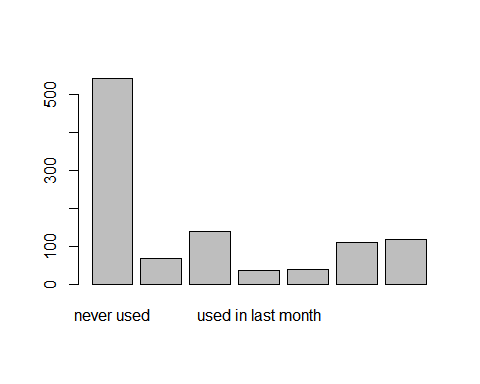
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$ethnicity)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$ethnicity): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$ethnicity  
## X-squared = 16.853, df = 6, p-value = 0.009839

Draw a plot of variable consumption\_amphetamines.

plot(drug\_train$consumption\_amphetamines)

 Label them for analysis.

drug\_train$consumption\_amphetamines <- factor(drug\_train$consumption\_amphetamines,  
 level=c("never used",  
 "used in last day",   
 "used in last week",  
 "used in last month",  
 "used in last year",  
 "used in last decade",  
 "used over a decade ago"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable consumption\_amphetamines.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_amphetamines)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 522 | 53 | 28 | 27 | 93 | 124 | 115 | 962 |   
## | 1.0 | 0.8 | 0.7 | 0.8 | 0.9 | 0.9 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 21 | 14 | 11 | 9 | 16 | 14 | 4 | 89 |   
## | 0.0 | 0.2 | 0.3 | 0.2 | 0.1 | 0.1 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 543 | 67 | 39 | 36 | 109 | 138 | 119 | 1051 |   
## | 0.5 | 0.1 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_amphetamines, workspace = 2e8)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_amphetamines  
## p-value = 6.015e-12  
## alternative hypothesis: two.sided

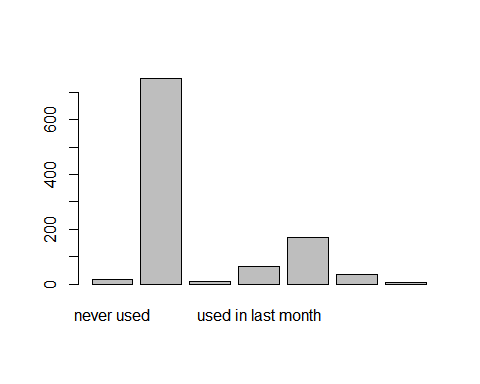
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_amphetamines)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$consumption\_amphetamines): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_amphetamines  
## X-squared = 70.401, df = 6, p-value = 3.383e-13

Draw a plot of variable consumption\_caffeine.

plot(drug\_train$consumption\_caffeine)

 Label them for analysis.

drug\_train$consumption\_caffeine <- factor(drug\_train$consumption\_caffeine,  
 level=c("never used",  
 "used in last day",   
 "used in last week",  
 "used in last month",  
 "used in last year",  
 "used in last decade",  
 "used over a decade ago"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable consumption\_caffeine.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_caffeine)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 15 | 691 | 148 | 57 | 34 | 11 | 6 | 962 |   
## | 0.9 | 0.9 | 0.9 | 0.9 | 1.0 | 1.0 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 1 | 59 | 22 | 7 | 0 | 0 | 0 | 89 |   
## | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 16 | 750 | 170 | 64 | 34 | 11 | 6 | 1051 |   
## | 0.0 | 0.7 | 0.2 | 0.1 | 0.0 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_caffeine, workspace = 2e8)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_caffeine  
## p-value = 0.1423  
## alternative hypothesis: two.sided

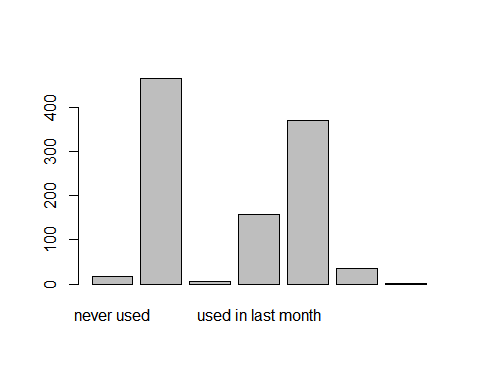
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_caffeine)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$consumption\_caffeine): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_caffeine  
## X-squared = 10.062, df = 6, p-value = 0.1221

Draw a plot of variable consumption\_chocolate.

plot(drug\_train$consumption\_chocolate)

 Label them for analysis.

drug\_train$consumption\_chocolate <- factor(drug\_train$consumption\_chocolate,  
 level=c("never used",  
 "used in last day",   
 "used in last week",  
 "used in last month",  
 "used in last year",  
 "used in last decade",  
 "used over a decade ago"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable consumption\_chocolate.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_chocolate)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 18 | 425 | 342 | 139 | 30 | 6 | 2 | 962 |   
## | 1.0 | 0.9 | 0.9 | 0.9 | 0.9 | 1.0 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 0 | 39 | 27 | 19 | 4 | 0 | 0 | 89 |   
## | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 18 | 464 | 369 | 158 | 34 | 6 | 2 | 1051 |   
## | 0.0 | 0.4 | 0.4 | 0.2 | 0.0 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_chocolate, workspace = 2e8)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_chocolate  
## p-value = 0.4581  
## alternative hypothesis: two.sided

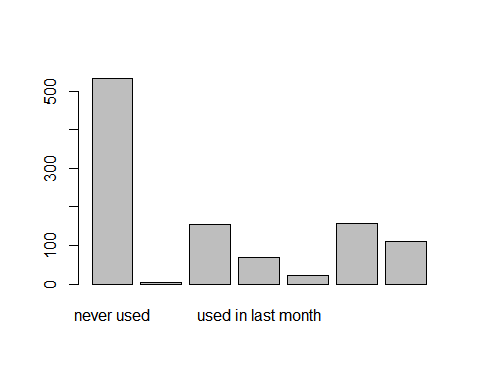
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_chocolate)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$consumption\_chocolate): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_chocolate  
## X-squared = 6.0946, df = 6, p-value = 0.4127

Draw a plot of variable consumption\_mushrooms.

plot(drug\_train$consumption\_mushrooms)

 Label them for analysis.

drug\_train$consumption\_mushrooms <- factor(drug\_train$consumption\_mushrooms,  
 level=c("never used",  
 "used in last day",   
 "used in last week",  
 "used in last month",  
 "used in last year",  
 "used in last decade",  
 "used over a decade ago"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable consumption\_mushrooms.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_mushrooms)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 512 | 3 | 20 | 57 | 129 | 133 | 108 | 962 |   
## | 1.0 | 1.0 | 0.9 | 0.8 | 0.8 | 0.9 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 21 | 0 | 3 | 12 | 28 | 22 | 3 | 89 |   
## | 0.0 | 0.0 | 0.1 | 0.2 | 0.2 | 0.1 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 533 | 3 | 23 | 69 | 157 | 155 | 111 | 1051 |   
## | 0.5 | 0.0 | 0.0 | 0.1 | 0.1 | 0.1 | 0.1 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_mushrooms, workspace = 2e8)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_mushrooms  
## p-value = 1.898e-09  
## alternative hypothesis: two.sided

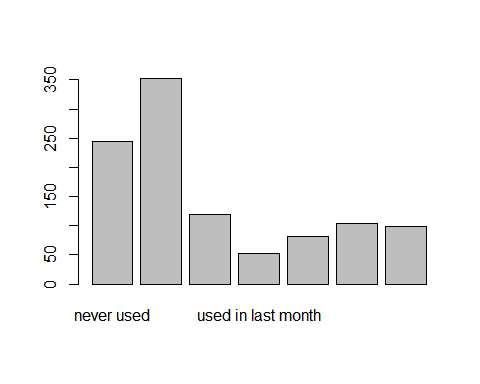
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_mushrooms)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$consumption\_mushrooms): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_mushrooms  
## X-squared = 51.171, df = 6, p-value = 2.736e-09

Draw a plot of variable consumption\_nicotine.

plot(drug\_train$consumption\_nicotine)

 Label them for analysis.

drug\_train$consumption\_nicotine <- factor(drug\_train$consumption\_nicotine,  
 level=c("never used",  
 "used in last day",   
 "used in last week",  
 "used in last month",  
 "used in last year",  
 "used in last decade",  
 "used over a decade ago"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable consumption\_nicotine.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_nicotine)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 237 | 293 | 74 | 50 | 97 | 115 | 96 | 962 |   
## | 1.0 | 0.8 | 0.9 | 0.9 | 0.9 | 1.0 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 7 | 59 | 8 | 3 | 6 | 4 | 2 | 89 |   
## | 0.0 | 0.2 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 244 | 352 | 82 | 53 | 103 | 119 | 98 | 1051 |   
## | 0.2 | 0.3 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_nicotine, workspace = 2e8)

##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_nicotine  
## p-value = 2.696e-09  
## alternative hypothesis: two.sided

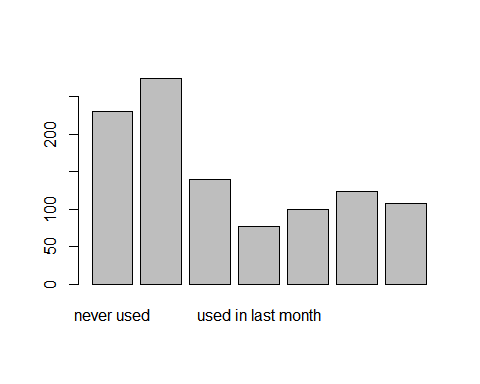
chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_nicotine)

## Warning in chisq.test(drug\_train$consumption\_cocaine\_last\_month,  
## drug\_train$consumption\_nicotine): Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_nicotine  
## X-squared = 51.973, df = 6, p-value = 1.888e-09

Draw a plot of variable consumption\_cannabis

plot(drug\_train$consumption\_cannabis)

 Label them for analysis.

drug\_train$consumption\_cannabis <- factor(drug\_train$consumption\_cannabis,  
 level=c("never used",  
 "used in last day",   
 "used in last week",  
 "used in last month",  
 "used in last year",  
 "used in last decade",  
 "used over a decade ago"),  
 labels = c(1:7))

Make contingency table and do Chi-square test and fisher test of variable consumption\_cannabis.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_cannabis)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Row Total |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## No | 226 | 234 | 80 | 70 | 113 | 132 | 107 | 962 |   
## | 1.0 | 0.9 | 0.8 | 0.9 | 0.9 | 0.9 | 1.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Yes | 4 | 40 | 20 | 7 | 10 | 7 | 1 | 89 |   
## | 0.0 | 0.1 | 0.2 | 0.1 | 0.1 | 0.1 | 0.0 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
## Column Total | 230 | 274 | 100 | 77 | 123 | 139 | 108 | 1051 |   
## | 0.2 | 0.3 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | |   
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_cannabis, workspace = 2e8)

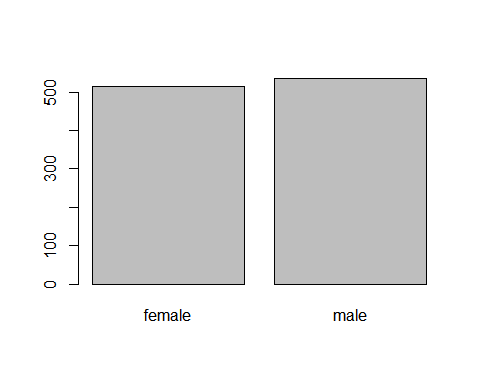
##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_cannabis  
## p-value = 1.13e-10  
## alternative hypothesis: two.sided

chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$consumption\_cannabis)

##   
## Pearson's Chi-squared test  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$consumption\_cannabis  
## X-squared = 53.973, df = 6, p-value = 7.469e-10

Draw a plot of variable gender.

plot(drug\_train$gender)

 Label them for analysis.

drug\_train$gender <- factor(drug\_train$gender,  
 level=c("female",  
 "male"),  
 labels = c(1:2))

Make contingency table and do Chi-square test and fisher test of variable gender.

get.contingency.table(drug\_train$consumption\_cocaine\_last\_month, drug\_train$gender)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | N / Col Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 1051   
##   
##   
## | indep.var   
## dep.var | 1 | 2 | Row Total |   
## -------------|-----------|-----------|-----------|  
## No | 459 | 503 | 962 |   
## | 0.9 | 0.9 | |   
## -------------|-----------|-----------|-----------|  
## Yes | 57 | 32 | 89 |   
## | 0.1 | 0.1 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 516 | 535 | 1051 |   
## | 0.5 | 0.5 | |   
## -------------|-----------|-----------|-----------|  
##   
##

fisher.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$gender, workspace = 2e8)

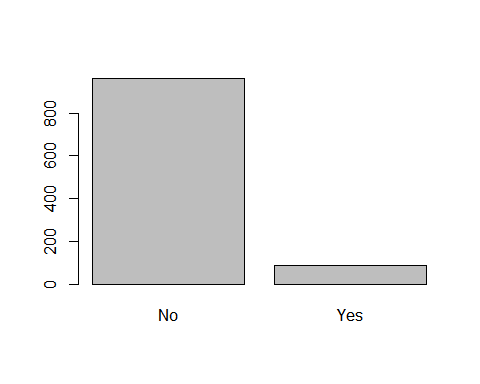
##   
## Fisher's Exact Test for Count Data  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$gender  
## p-value = 0.003792  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.3155844 0.8204033  
## sample estimates:  
## odds ratio   
## 0.5126218

chisq.test(drug\_train$consumption\_cocaine\_last\_month, drug\_train$gender)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: drug\_train$consumption\_cocaine\_last\_month and drug\_train$gender  
## X-squared = 8.0531, df = 1, p-value = 0.004543

Draw a plot of variable consumption\_cocaine\_last\_month

plot(drug\_train$consumption\_cocaine\_last\_month)

 Label them for analysis.

drug\_train$consumption\_cocaine\_last\_month <- factor(drug\_train$consumption\_cocaine\_last\_month,  
 level=c("Yes",  
 "No"),  
 labels = c(0:1))

For numeric variables, i used t-test to check the significance.

Do t-test of variable personality\_neuroticism.

t.test(drug\_train$personality\_neuroticism ~ drug\_train$consumption\_cocaine\_last\_month)

##   
## Welch Two Sample t-test  
##   
## data: drug\_train$personality\_neuroticism by drug\_train$consumption\_cocaine\_last\_month  
## t = 2.5582, df = 108, p-value = 0.01191  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## 0.8884619 7.0033132  
## sample estimates:  
## mean in group 0 mean in group 1   
## 55.07978 51.13389

Do t-test of variable personality\_extraversion.

t.test(drug\_train$personality\_extraversion ~ drug\_train$consumption\_cocaine\_last\_month)

##   
## Welch Two Sample t-test  
##   
## data: drug\_train$personality\_extraversion by drug\_train$consumption\_cocaine\_last\_month  
## t = 1.5197, df = 97.753, p-value = 0.1318  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -0.9612645 7.2458962  
## sample estimates:  
## mean in group 0 mean in group 1   
## 52.73483 49.59252

Do t-test of variable personality\_openness.

t.test(drug\_train$personality\_openness ~ drug\_train$consumption\_cocaine\_last\_month)

##   
## Welch Two Sample t-test  
##   
## data: drug\_train$personality\_openness by drug\_train$consumption\_cocaine\_last\_month  
## t = 1.8766, df = 105.2, p-value = 0.06334  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -0.1899466 6.9046141  
## sample estimates:  
## mean in group 0 mean in group 1   
## 56.44382 53.08649

Do t-test of variable personality\_agreeableness.

t.test(drug\_train$personality\_agreeableness ~ drug\_train$consumption\_cocaine\_last\_month)

##   
## Welch Two Sample t-test  
##   
## data: drug\_train$personality\_agreeableness by drug\_train$consumption\_cocaine\_last\_month  
## t = -2.6375, df = 102.63, p-value = 0.009649  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -7.665793 -1.085280  
## sample estimates:  
## mean in group 0 mean in group 1   
## 45.89213 50.26767

Do t-test of variable personality\_conscientiousness.

t.test(drug\_train$personality\_conscientiousness ~ drug\_train$consumption\_cocaine\_last\_month)

##   
## Welch Two Sample t-test  
##   
## data: drug\_train$personality\_conscientiousness by drug\_train$consumption\_cocaine\_last\_month  
## t = -3.5359, df = 108.24, p-value = 0.0005994  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## -8.237717 -2.319644  
## sample estimates:  
## mean in group 0 mean in group 1   
## 45.21124 50.48992

Do t-test of variable personality\_impulsiveness.

t.test(drug\_train$personality\_impulsiveness ~ drug\_train$consumption\_cocaine\_last\_month)

##   
## Welch Two Sample t-test  
##   
## data: drug\_train$personality\_impulsiveness by drug\_train$consumption\_cocaine\_last\_month  
## t = 4.9558, df = 107.13, p-value = 2.711e-06  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## 5.365992 12.520877  
## sample estimates:  
## mean in group 0 mean in group 1   
## 55.63034 46.68690

Do t-test of variable personality\_sensation.

t.test(drug\_train$personality\_sensation ~ drug\_train$consumption\_cocaine\_last\_month)

##   
## Welch Two Sample t-test  
##   
## data: drug\_train$personality\_sensation by drug\_train$consumption\_cocaine\_last\_month  
## t = 7.0187, df = 105.43, p-value = 2.251e-10  
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0  
## 95 percent confidence interval:  
## 12.76762 22.82111  
## sample estimates:  
## mean in group 0 mean in group 1   
## 68.99551 51.20114

After these test, we chose below variables: “age”, “gender”, “country”, “ethnicity”, “consumption\_alcohol”,“consumption\_amphetamines”,“consumption\_cannabis”,“consumption\_mushrooms”,“consumption\_nicotine”, “consumption\_cocaine\_last\_month”,“personality\_neuroticism”, “personality\_agreeableness”,“personality\_conscientiousness”, “personality\_impulsiveness”, “personality\_sensation”

Creat a new dataframe with selected variables.

load("C:/Users/zhaoz/Desktop/jiqixuexi/pro/drug\_train\_test.RData")  
drug\_train\_pro <- drug\_train[,c("age", "gender", "country", "ethnicity", "consumption\_alcohol",   
 "consumption\_amphetamines","consumption\_cannabis",  
 "consumption\_mushrooms","consumption\_nicotine", "consumption\_cocaine\_last\_month",  
 "personality\_neuroticism", "personality\_agreeableness",   
 "personality\_conscientiousness", "personality\_impulsiveness", "personality\_sensation")]

And for testing sample.

drug\_test <- drug\_test[,c("age", "gender", "country", "ethnicity", "consumption\_alcohol",   
 "consumption\_amphetamines","consumption\_cannabis",  
 "consumption\_mushrooms","consumption\_nicotine", "consumption\_cocaine\_last\_month",  
 "personality\_neuroticism", "personality\_agreeableness",   
 "personality\_conscientiousness", "personality\_impulsiveness", "personality\_sensation")]

## KNN model

Firstly, we use drug\_train\_pro this new dataset to check the performance of the model.

Define the training control, I use 5-fold cross validation.

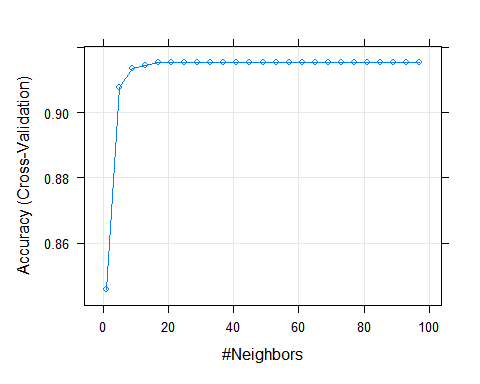
## set “k”s  
different\_k <- data.frame(k = seq(1, 99, 4))  
## use 5-fold cross validation  
ctrl\_cv5 <- trainControl(method = "cv",  
 number = 5)

## Model 1

## run the training  
set.seed(2222)  
  
drug\_train\_pro\_knn\_tuned <-   
 train(consumption\_cocaine\_last\_month ~ .,   
 data = drug\_train\_pro,  
 method = "knn",  
 # validation used!  
 trControl = ctrl\_cv5,  
 # parameters to be compared  
 tuneGrid = different\_k)  
  
## now validation is applied to EVERY SINGLE value of k.  
  
## lets check the results  
  
drug\_train\_pro\_knn\_tuned

## k-Nearest Neighbors   
##   
## 1051 samples  
## 14 predictor  
## 2 classes: 'No', 'Yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 841, 841, 841, 840, 841   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 1 0.8458768 0.037934772  
## 5 0.9077003 0.035178529  
## 9 0.9134146 0.014461653  
## 13 0.9143670 -0.001820941  
## 17 0.9153193 0.000000000  
## 21 0.9153193 0.000000000  
## 25 0.9153193 0.000000000  
## 29 0.9153193 0.000000000  
## 33 0.9153193 0.000000000  
## 37 0.9153193 0.000000000  
## 41 0.9153193 0.000000000  
## 45 0.9153193 0.000000000  
## 49 0.9153193 0.000000000  
## 53 0.9153193 0.000000000  
## 57 0.9153193 0.000000000  
## 61 0.9153193 0.000000000  
## 65 0.9153193 0.000000000  
## 69 0.9153193 0.000000000  
## 73 0.9153193 0.000000000  
## 77 0.9153193 0.000000000  
## 81 0.9153193 0.000000000  
## 85 0.9153193 0.000000000  
## 89 0.9153193 0.000000000  
## 93 0.9153193 0.000000000  
## 97 0.9153193 0.000000000  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 97.

plot(drug\_train\_pro\_knn\_tuned)



We can see that from the perspective of AUC the best value of k = 97. Value of Accuracy is good, the best accuracy obtained >> k = 97 and the accuracy seems to increase with k.

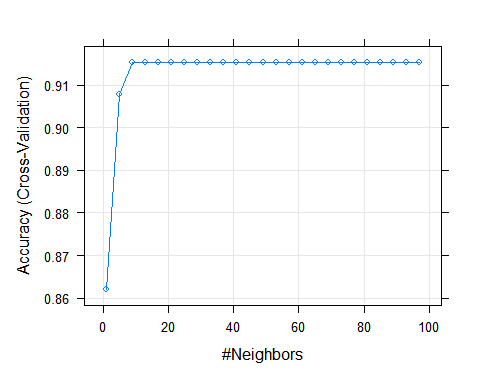
## Model 2

In the case of cross-validation, should apply transformations of input data in the train() function, I use preProcess = “range”.

set.seed(222)  
  
drug\_train\_pro\_knn\_tuned\_scaled <-   
 train(consumption\_cocaine\_last\_month ~ .,   
 data = drug\_train\_pro,  
 method = "knn",  
 # validation used!  
 trControl = ctrl\_cv5,  
 # parameters to be compared  
 tuneGrid = different\_k,  
 # data transformation  
 preProcess = c("range"))  
  
  
drug\_train\_pro\_knn\_tuned\_scaled

## k-Nearest Neighbors   
##   
## 1051 samples  
## 14 predictor  
## 2 classes: 'No', 'Yes'   
##   
## Pre-processing: re-scaling to [0, 1] (53)   
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 840, 841, 841, 841, 841   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 1 0.8620447 0.07447009  
## 5 0.9077138 0.01800999  
## 9 0.9153193 0.03564817  
## 13 0.9153193 0.00000000  
## 17 0.9153193 0.00000000  
## 21 0.9153193 0.00000000  
## 25 0.9153193 0.00000000  
## 29 0.9153193 0.00000000  
## 33 0.9153193 0.00000000  
## 37 0.9153193 0.00000000  
## 41 0.9153193 0.00000000  
## 45 0.9153193 0.00000000  
## 49 0.9153193 0.00000000  
## 53 0.9153193 0.00000000  
## 57 0.9153193 0.00000000  
## 61 0.9153193 0.00000000  
## 65 0.9153193 0.00000000  
## 69 0.9153193 0.00000000  
## 73 0.9153193 0.00000000  
## 77 0.9153193 0.00000000  
## 81 0.9153193 0.00000000  
## 85 0.9153193 0.00000000  
## 89 0.9153193 0.00000000  
## 93 0.9153193 0.00000000  
## 97 0.9153193 0.00000000  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 97.

plot(drug\_train\_pro\_knn\_tuned\_scaled)

 We can see that from the perspective of AUC the best value of k = 9. Value of Accuracy is very similar to above.

Check which value of k gives the highest AUC.

ctrl\_cv5a <- trainControl(method = "cv",  
 number = 5,  
 # probabilities of each level predicted in cross-validation  
 classProbs = TRUE,  
 # summary function that includes ROC  
 summaryFunction = twoClassSummary)

## Model 3

In the case of cross-validation, should apply transformations of input data in the train() function, I use preProcess = “range”. And I also use validation with probabilities and twoClassSummary

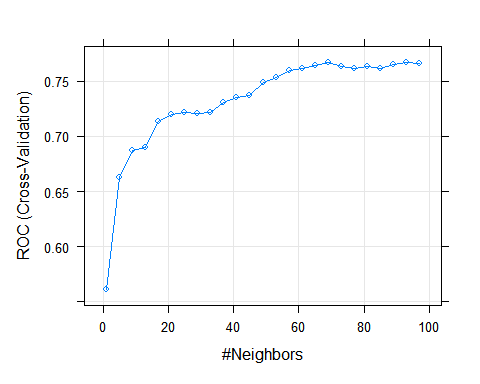
set.seed(2222)  
drug\_train\_pro\_knn\_tuned\_scaled2 <-   
 train(consumption\_cocaine\_last\_month ~ .,   
 data = drug\_train\_pro,  
 # knn  
 method = "knn",  
 # validation used - with probabilities and twoClassSummary  
 trControl = ctrl\_cv5a,  
 # parameters to be compared  
 tuneGrid = different\_k,  
 preProcess = c("range"),  
 metric = "ROC")

## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day  
  
## Warning in preProcess.default(thresh = 0.95, k = 5, freqCut = 19, uniqueCut =  
## 10, : No variation for for: consumption\_mushroomsused in last day

drug\_train\_pro\_knn\_tuned\_scaled2

## k-Nearest Neighbors   
##   
## 1051 samples  
## 14 predictor  
## 2 classes: 'No', 'Yes'   
##   
## Pre-processing: re-scaling to [0, 1] (53)   
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 841, 841, 841, 840, 841   
## Resampling results across tuning parameters:  
##   
## k ROC Sens Spec   
## 1 0.5612582 0.9303595 0.19215686  
## 5 0.6628637 0.9844182 0.02287582  
## 9 0.6872715 0.9989637 0.01111111  
## 13 0.6899271 1.0000000 0.00000000  
## 17 0.7134254 1.0000000 0.00000000  
## 21 0.7200541 1.0000000 0.00000000  
## 25 0.7214100 1.0000000 0.00000000  
## 29 0.7204547 1.0000000 0.00000000  
## 33 0.7212708 1.0000000 0.00000000  
## 37 0.7307262 1.0000000 0.00000000  
## 41 0.7350656 1.0000000 0.00000000  
## 45 0.7374007 1.0000000 0.00000000  
## 49 0.7491680 1.0000000 0.00000000  
## 53 0.7533764 1.0000000 0.00000000  
## 57 0.7594637 1.0000000 0.00000000  
## 61 0.7615735 1.0000000 0.00000000  
## 65 0.7645632 1.0000000 0.00000000  
## 69 0.7669268 1.0000000 0.00000000  
## 73 0.7636343 1.0000000 0.00000000  
## 77 0.7618099 1.0000000 0.00000000  
## 81 0.7632656 1.0000000 0.00000000  
## 85 0.7613107 1.0000000 0.00000000  
## 89 0.7655147 1.0000000 0.00000000  
## 93 0.7668845 1.0000000 0.00000000  
## 97 0.7662503 1.0000000 0.00000000  
##   
## ROC was used to select the optimal model using the largest value.  
## The final value used for the model was k = 69.

plot(drug\_train\_pro\_knn\_tuned\_scaled2)



Now run the model for testing sample and check. Testing sample is names “drug\_test”, separated from “drug\_train”.

drug\_test$predicted <- drug\_train\_pro\_knn\_tuned\_scaled2 %>% predict(drug\_test)

Lets built a confusion matrix table.

# confusion matrix table   
confusion\_table = table(drug\_test[,'predicted'], drug\_test[,'consumption\_cocaine\_last\_month'])  
confusion\_table

##   
## No Yes  
## No 411 38  
## Yes 0 0

Definite of TP,TN,FP,FN.

## Definition of TP,TN,FP,FN  
confusion\_table[1,1] = 'TN'  
confusion\_table[1,2] = 'FN'  
confusion\_table[2,1] = 'FP'  
confusion\_table[2,2] = 'TP'  
## output  
confusion\_table

##   
## No Yes  
## No TN FN   
## Yes FP TP

Write a function to calculate accuracy.

## function  
et\_accuracy <- function(df, predicted, actual){  
 confusion\_table = table(drug\_test[,'predicted'], drug\_test[,'consumption\_cocaine\_last\_month'])  
 TP = confusion\_table[2,2]  
 TN = confusion\_table[1,1]  
 FN = confusion\_table[1,2]  
 FP = confusion\_table[2,1]  
 accuracy = round((TP + TN) / sum(TP,FP,TN,FN), 2)  
 return(accuracy)  
}

Use this function to check the accuracy of testing sample!

## accuracy value   
score = data.frame(accuracy=et\_accuracy(drug\_test,'predicted','consumption\_cocaine\_last\_month'))  
  
## output  
score

## accuracy  
## 1 0.92

Seems good!

## Conclusion

I also used SVM and decision tree, the best method is KNN, because the Accuracy value is highest.