L10

Kovel

1 РЅРѕСЏР±СЂСЏ 2020 Рі

# Download the data

set.seed(123)  
path='C:/education.csv'

## Висновок: для побудови моделі кластеризації використано дані про учнів школ двох типів. Залишимо для моделювання лише кількісні змінні. Датасет не потребує розподілу на навчальну та тестову вибірки, не потребує шкалювання.

f <- read.csv(file = path, header = TRUE, encoding = 'UNICOD')  
head(f)

## school sex age reason guardian traveltime studytime activities internet  
## 1 GP F 18 course mother 2 2 no no  
## 2 GP F 17 course father 1 2 no yes  
## 3 GP F 15 other mother 1 2 no yes  
## 4 GP F 15 home mother 1 3 yes yes  
## 5 GP F 16 home father 1 2 no no  
## 6 GP M 16 reputation mother 1 2 yes yes

## Загружені данні

## Factors as numeric

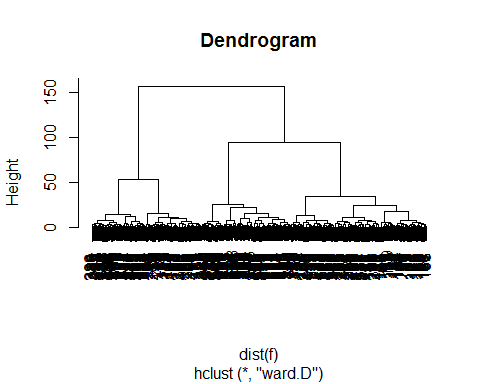
f$school <- as.numeric(as.factor(f$school))-1  
f$sex <- as.numeric(as.factor(f$sex))-1  
f$reason <- as.numeric(as.factor(f$reason))-1  
f$guardian <- as.numeric(as.factor(f$guardian))-1  
f$activities <- as.numeric(as.factor(f$activities))-1  
f$internet <- as.numeric(as.factor(f$internet))-1  
head (f)

## school sex age reason guardian traveltime studytime activities internet  
## 1 0 0 18 0 1 2 2 0 0  
## 2 0 0 17 0 0 1 2 0 1  
## 3 0 0 15 2 1 1 2 0 1  
## 4 0 0 15 1 1 1 3 1 1  
## 5 0 0 16 1 0 1 2 0 0  
## 6 0 1 16 3 1 1 2 1 1

## Качествен данні переведені у кількісні

# Hierarchical clustering

model\_hc <- hclust(dist(f), method = "ward.D" )  
plot(model\_hc, main = paste('Dendrogram'))

 ##Висновок: на основі навчальної вибірки побудовано дендрограму з використанням методу Ward.D. ## Fitting HC to the dataset

y\_hc <- cutree(model\_hc, k = 3)  
#cluster cores  
aggregate(f,by=list(y\_hc),FUN=mean)

## Group.1 school sex age reason guardian traveltime  
## 1 1 0.33333333 0.4761905 18.00952 0.6095238 0.9619048 1.609524  
## 2 2 0.03125000 0.5375000 15.95000 0.4312500 0.7750000 1.412500  
## 3 3 0.04615385 0.3923077 16.55385 2.7923077 0.8615385 1.361538  
## studytime activities internet  
## 1 2.057143 0.4190476 0.7714286  
## 2 1.868750 0.5062500 0.8625000  
## 3 2.223077 0.5846154 0.8461538

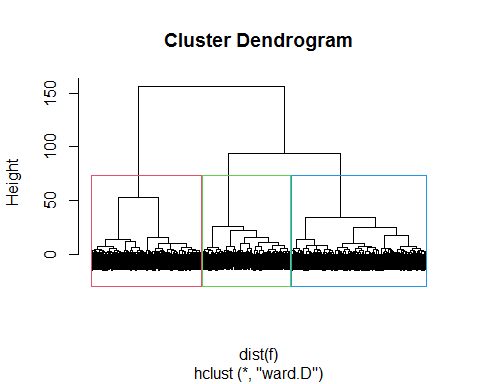
#Cluster stat  
f$hc <- y\_hc  
table(f$hc)

##   
## 1 2 3   
## 105 160 130

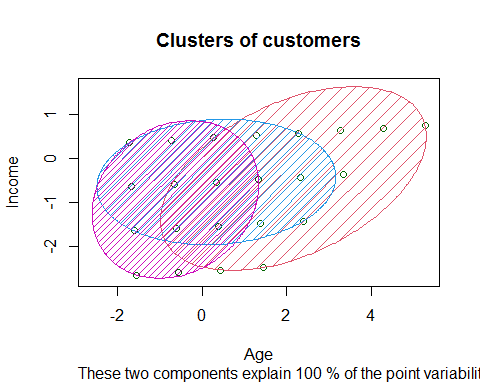
## Висновок: на основі аналізу дендрограми виявлено три кластера: 1 – “дорослі”, 2 – “ малого віку ”, 3 – “середнього віку ”. Розраховано характеристики типового об’єкту кластерів.

## Plotting the dendrogram

plot(model\_hc, cex = 0.7, labels = FALSE)  
rect.hclust(model\_hc, k = 3, border = 2:5)

 ##Висновок: проведено візуалізацію кластерів на дендрограмі. ## Visualising the clusters

library(cluster)  
clusplot(f[,c('age','traveltime')],  
 y\_hc,  
 lines = 0,  
 shade = TRUE,  
 color = TRUE,  
 labels= 0,  
 plotchar = FALSE,  
 span = TRUE,  
 main = paste('Clusters of customers'),  
 xlab = 'Age',  
 ylab = 'Income')

 ##Висновок: проведено візуалізацію кластерів на датасеті.