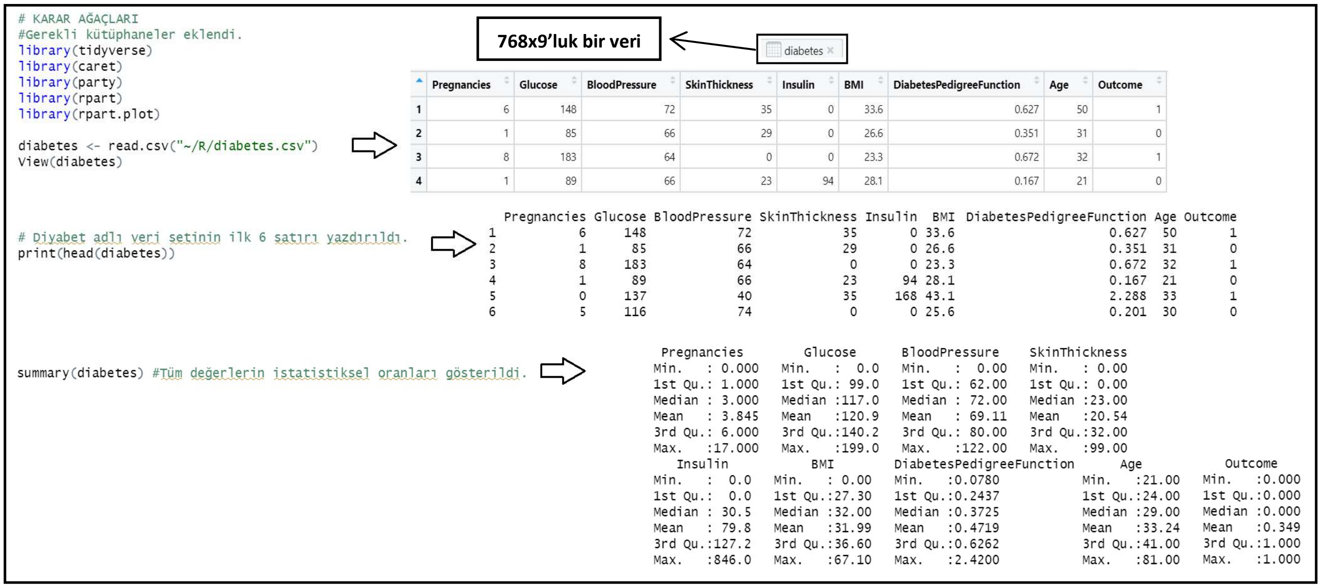
# VERİ MADENCİLİĞİ UYGULAMASI

**Diyabet hastalığına ait karar ağaçları ile ilgili kodlar aşağıda görüldüğü gibidir.**

## Karar ağaçlarına ait kod çıktıları:



```
# KARAR AĞAÇLARI
#Gerekli kütüphaneler eklendi.
library(tidyverse)
library(caret)
library(party)
library(rpart)
library(rpart.plot)

diabetes <- read.csv("~/R/diabetes.csv")
View(diabetes)
```

768x9'luk bir veri

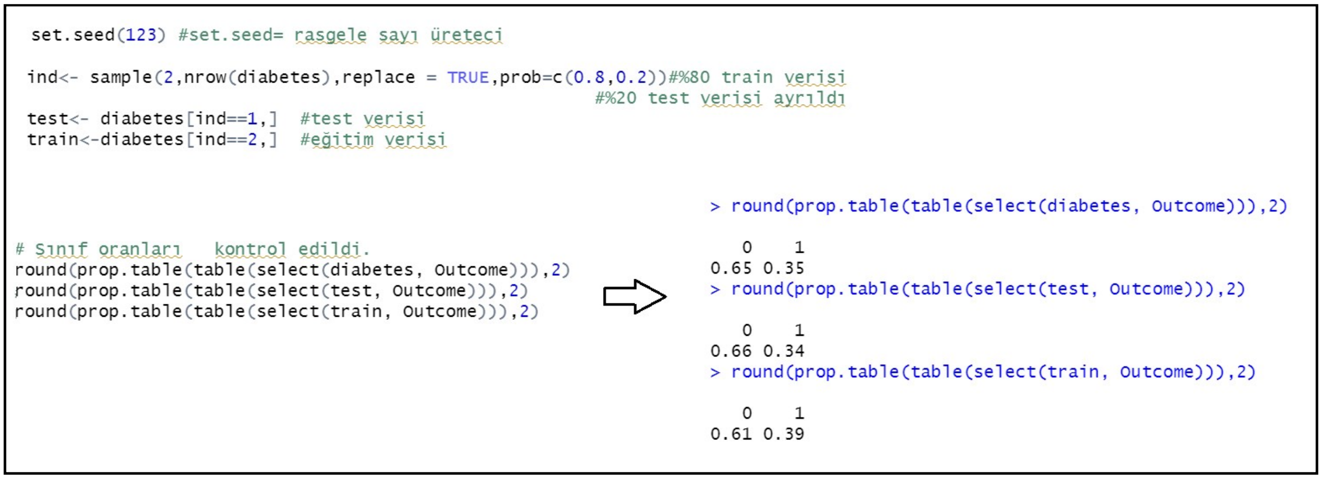
| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 2 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 3 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 4 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |

```
# Diyabet adlı veri setinin ilk 6 satırı yazdırıldı.
print(head(diabetes))
```

```
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome
1           6     148            72            35       0 33.6                    0.627  50       1
2           1      85            66            29       0 26.6                    0.351  31       0
3           8     183            64             0       0 23.3                    0.672  32       1
4           1      89            66            23      94 28.1                    0.167  21       0
5           0     137            40            35     168 43.1                    2.288  33       1
6           5     116            74             0       0 25.6                    0.201  30       0
```

```
summary(diabetes) #Tüm değerlerin istatistiksel oranları gösterildi.
```

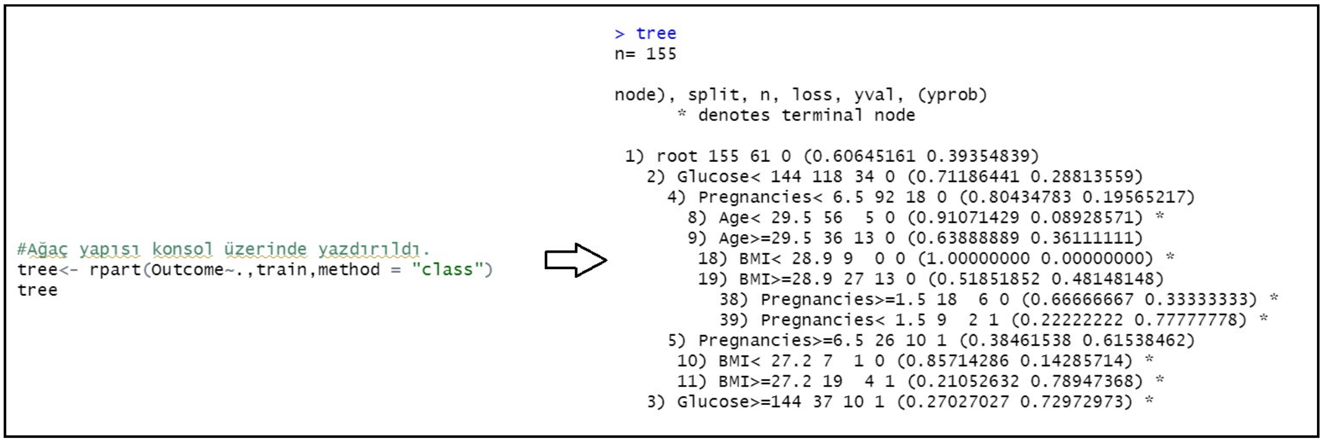
```
  Pregnancies        Glucose       BloodPressure    SkinThickness
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    Insulin           BMI        DiabetesPedigreeFunction      Age           Outcome
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00   Min.   :0.000
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
 Median : 30.5   Median :32.00   Median :0.3725           Median :29.00   Median :0.000
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24   Mean   :0.349
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
 Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00   Max.   :1.000
```



```
 set.seed(123) #set.seed= rasgele sayı üreteci

ind<- sample(2,nrow(diabetes),replace = TRUE,prob=c(0.8,0.2))#%80 train verisi
                                                             #%20 test verisi ayrıldı

test<- diabetes[ind==1,]  #test verisi
train<-diabetes[ind==2,]  #eğitim verisi
```

```
# Sınıf oranları   kontrol edildi.
round(prop.table(table(select(diabetes, Outcome))),2)
round(prop.table(table(select(test, Outcome))),2)
round(prop.table(table(select(train, Outcome))),2)
```

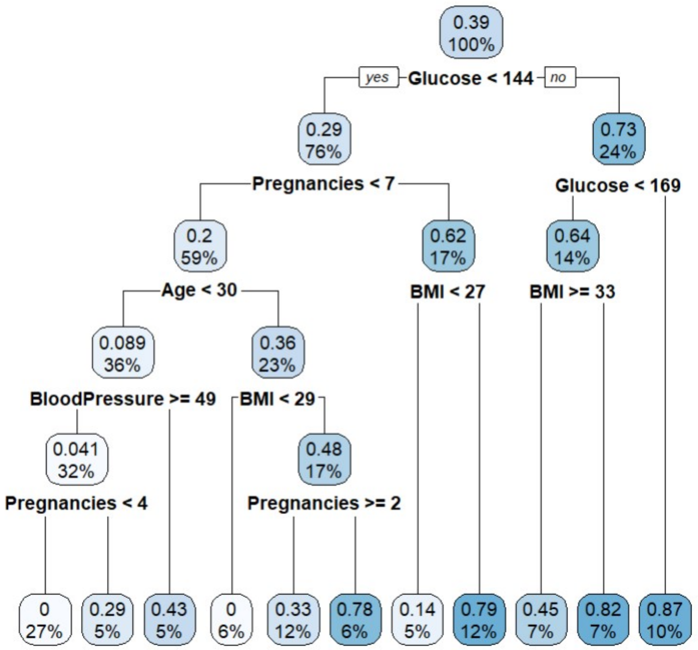
```
> round(prop.table(table(select(diabetes, Outcome))),2)

   0    1 
0.65 0.35 
> round(prop.table(table(select(test, Outcome))),2)

   0    1 
0.66 0.34 
> round(prop.table(table(select(train, Outcome))),2)

   0    1 
0.61 0.39 
```



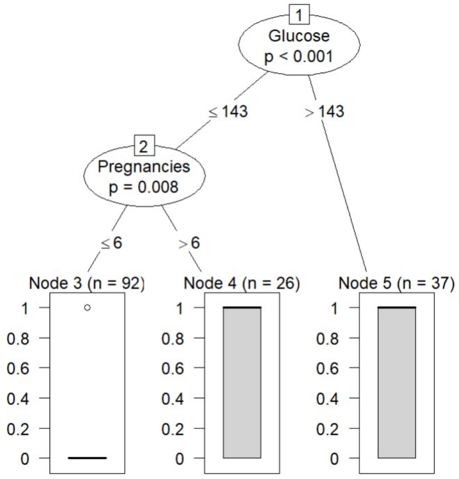
```
#Ağaç yapısı konsol üzerinde yazdırıldı.
tree<- rpart(Outcome~.,train,method = "class")
tree
```

```
> tree
n= 155 

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 155 61 0 (0.60645161 0.39354839)  
   2) Glucose< 144 118 34 0 (0.71186441 0.28813559)  
     4) Pregnancies< 6.5 92 18 0 (0.80434783 0.19565217)  
       8) Age< 29.5 56  5 0 (0.91071429 0.08928571) *
       9) Age>=29.5 36 13 0 (0.63888889 0.36111111)  
        18) BMI< 28.9 9  0 0 (1.00000000 0.00000000) *
        19) BMI>=28.9 27 13 0 (0.51851852 0.48148148)  
          38) Pregnancies>=1.5 18  6 0 (0.66666667 0.33333333) *
          39) Pregnancies< 1.5 9  2 1 (0.22222222 0.77777778) *
     5) Pregnancies>=6.5 26 10 1 (0.38461538 0.61538462)  
      10) BMI< 27.2 7  1 0 (0.85714286 0.14285714) *
      11) BMI>=27.2 19  4 1 (0.21052632 0.78947368) *
   3) Glucose>=144 37 10 1 (0.27027027 0.72972973) *
```

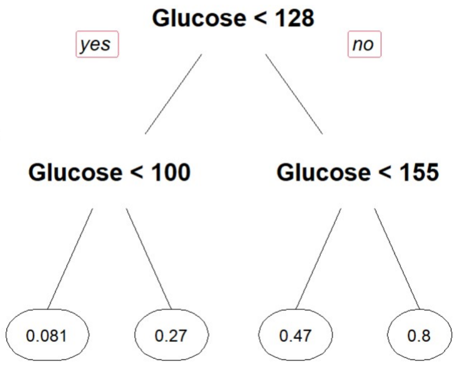
```
#Ağaç yapısı çizdirildi
tree<- rpart(Outcome~.,train)
rpart.plot(tree)
```

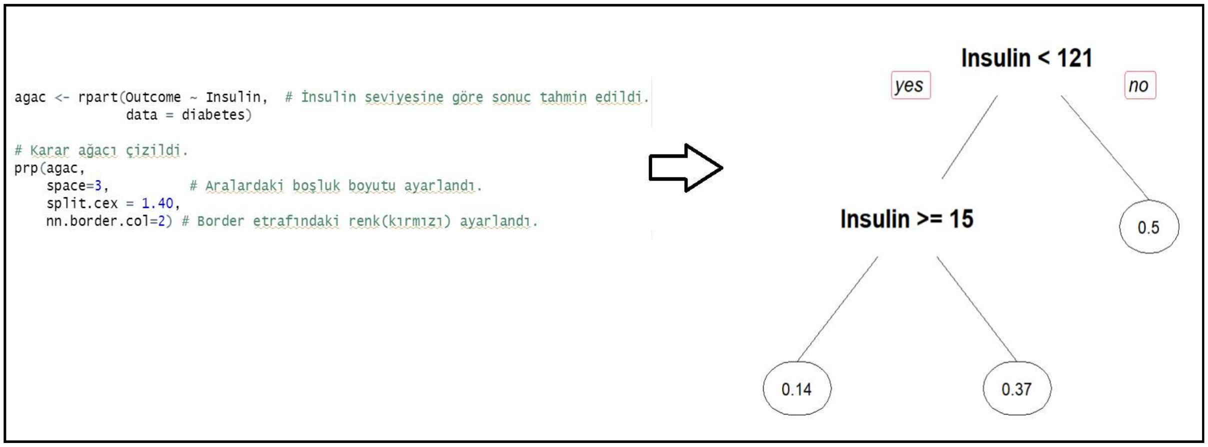


```
#ctree( Koşullu Çıkarım Ağaçları)fonksiyonu ile karar ağacı yapısı çizdirildi.)
model<- ctree(Outcome ~ ., train)
plot(model)
```

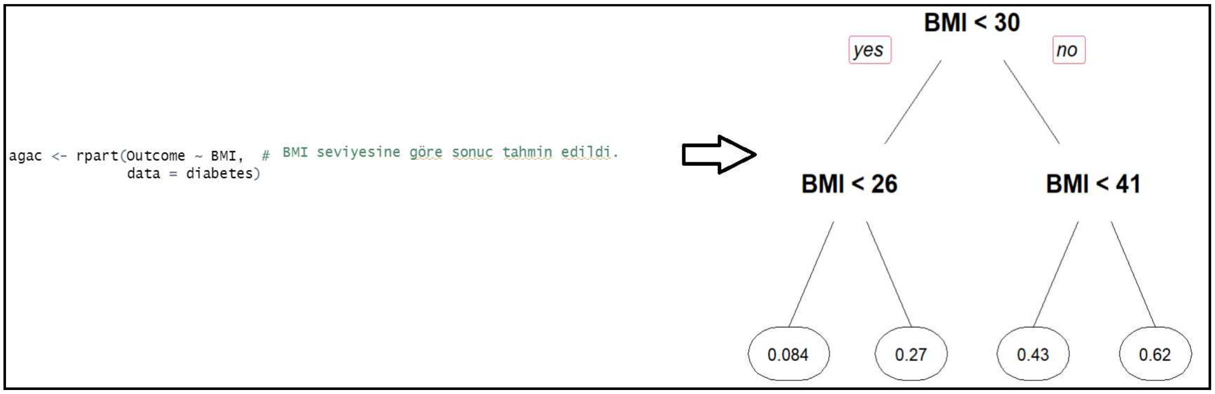


```
######
agac <- rpart(Outcome ~ Glucose,          # Glukoza göre sonuc tahmin edildi.
                   data = diabetes)       # Diyabet adlı veri setine eşitlendi.
```

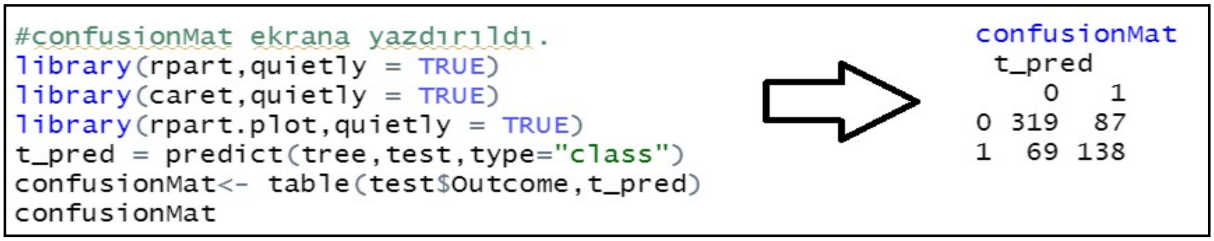




```r
agac <- rpart(Outcome ~ Insulin,  # İnsulin seviyesine göre sonuc tahmin edildi.
              data = diabetes)

# Karar ağacı çizildi.
prp(agac,
    space=3,          # Aralardaki boşluk boyutu ayarlandı.
    split.cex = 1.40,
    nn.border.col=2) # Border etrafındaki renk(kırmızı) ayarlandı.
```

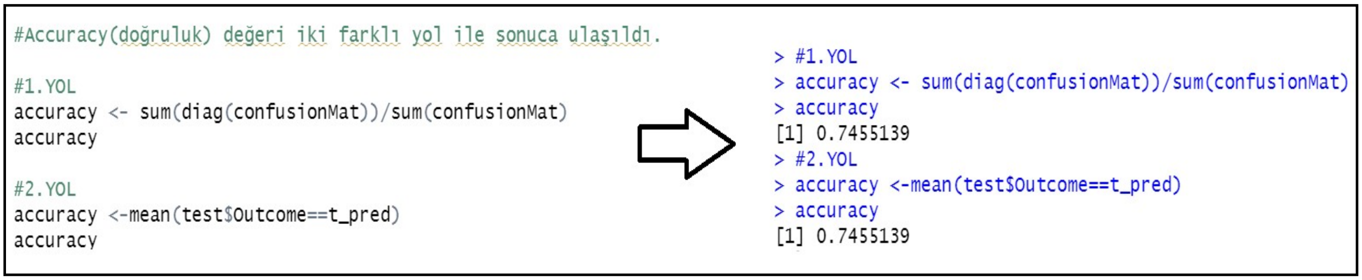


```r
agac <- rpart(Outcome ~ BMI,  #  BMI seviyesine göre sonuc tahmin edildi.
              data = diabetes)
```



```r
#confusionMat ekrana yazdırıldı.
library(rpart,quietly = TRUE)
library(caret,quietly = TRUE)
library(rpart.plot,quietly = TRUE)
t_pred = predict(tree,test,type="class")
confusionMat<- table(test$Outcome,t_pred)
confusionMat
```

```
confusionMat
  t_pred
      0   1
0   319  87
1    69 138
```

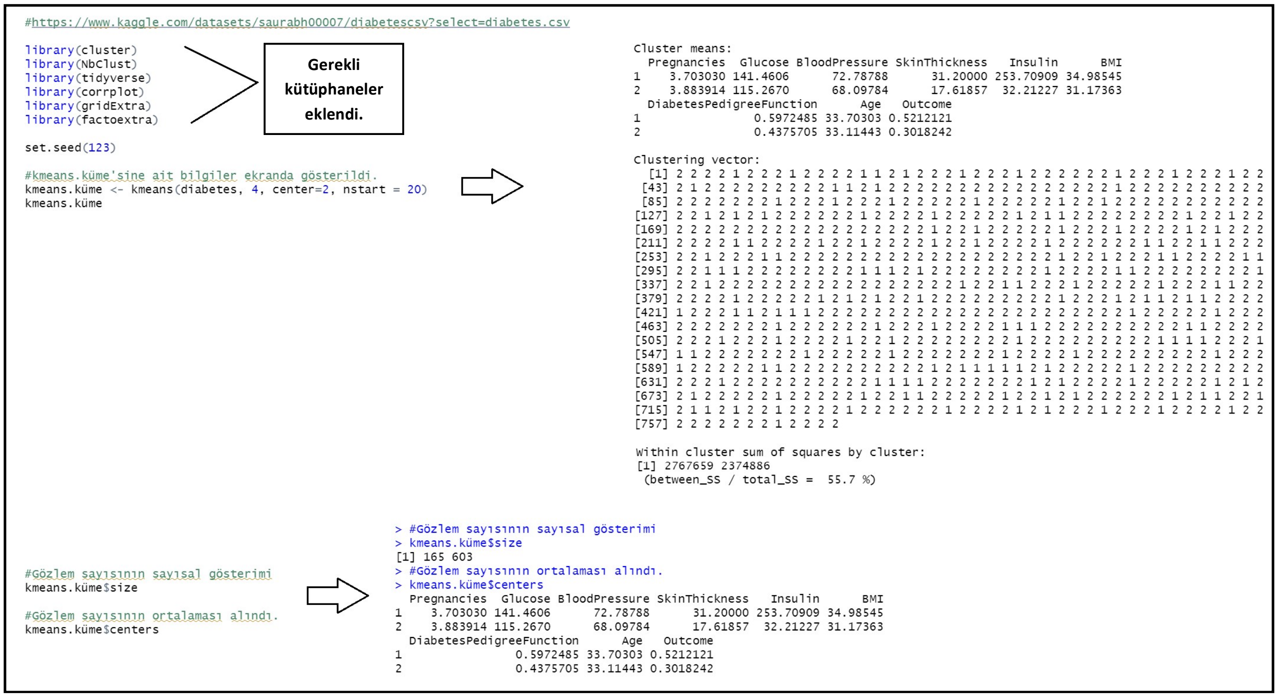


```r
#Accuracy(doğruluk) değeri iki farklı yol ile sonuca ulaşıldı.

#1.YOL
accuracy <- sum(diag(confusionMat))/sum(confusionMat)
accuracy

#2.YOL
accuracy <-mean(test$Outcome==t_pred)
accuracy
```

```
> #1.YOL
> accuracy <- sum(diag(confusionMat))/sum(confusionMat)
> accuracy
[1] 0.7455139
> #2.YOL
> accuracy <-mean(test$Outcome==t_pred)
> accuracy
[1] 0.7455139
```

DEVAMI

## Kümelemeye ait kod çıktıları:



```
#https://www.kaggle.com/datasets/saurabh00007/diabetescsv?select=diabetes.csv

library(cluster)
library(NbClust)
library(tidyverse)
library(corrplot)
library(gridExtra)
library(factoextra)

set.seed(123)

#kmeans.küme'sine ait bilgiler ekranda gösterildi.
kmeans.küme <- kmeans(diabetes, 4, center=2, nstart = 20)
kmeans.küme
```

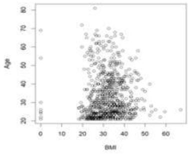
Gerekli kütüphaneler eklendi.

```
Cluster means:
  Pregnancies  Glucose BloodPressure SkinThickness   Insulin      BMI
1    3.703030 141.4606      72.78788      31.20000 253.70909 34.98545
2    3.883914 115.2670      68.09784      17.61857  32.21227 31.17363
  DiabetesPedigreeFunction      Age   Outcome
1                0.5972485 33.70303 0.5212121
2                0.4375705 33.11443 0.3018242

Clustering vector:
  [1] 2 2 2 2 1 2 2 2 1 2 2 2 2 1 1 2 1 2 2 2 1 2 2 2 1 2 2 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2
 [43] 2 1 2 2 2 2 2 2 2 2 2 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2
 [85] 2 2 2 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2 2 2 2
[127] 2 2 1 2 1 2 1 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2 1 2 1 1 2 2 2 2 2 2 2 2 1 2 2 1 2 2
[169] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 1 2 2 2 1 2 2 2 2 1 2 1 2 2 2
[211] 2 2 2 2 1 1 2 2 2 2 1 2 2 1 2 2 2 2 1 2 2 1 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2 1 1 2 2 2
[253] 2 2 1 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 1 1 2 2 2 2 2 1 1
[295] 2 2 1 1 1 2 2 2 2 2 2 2 2 1 1 1 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 1 2 2 2 2 2 2 2 2 1
[337] 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 1 2 2 2 1 2 2 2 2 2 1 2 2 2 1 1 2 2
[379] 2 2 2 2 1 2 2 2 2 2 1 2 1 2 1 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 1 2 1 1 2 1 1 2 2 2 2 2
[421] 1 2 2 2 1 1 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2
[463] 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 1 2 2 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2
[505] 2 2 2 1 2 2 2 1 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2 1
[547] 1 1 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2
[589] 1 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 1 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2
[631] 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 1 1 2 2 2 2 2 2 2 1 2 1 2 2 2 2 1 2 2 2 2 2 1 2 1 2
[673] 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 2 1 1 2 2 2 2 2 1 2 2 1 2 1 2 2 2 2 2 2 1 2 1 1 2 2 1
[715] 2 1 1 2 1 2 2 1 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 1 2 1 2 2 2 1 2 2 2 1 2 2 2 2 1 2 2
[757] 2 2 2 2 2 2 2 1 2 2 2 2

Within cluster sum of squares by cluster:
[1] 2767659 2374886
 (between_SS / total_SS =  55.7 %)
```

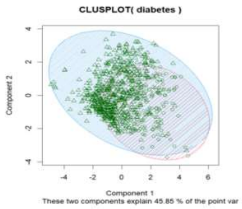
```
#Gözlem sayısının sayısal gösterimi
kmeans.küme$size

#Gözlem sayısının ortalaması alındı.
kmeans.küme$centers
```

```
> #Gözlem sayısının sayısal gösterimi
> kmeans.küme$size
[1] 165 603
> #Gözlem sayısının ortalaması alındı.
> kmeans.küme$centers
  Pregnancies  Glucose BloodPressure SkinThickness   Insulin      BMI
1    3.703030 141.4606      72.78788      31.20000 253.70909 34.98545
2    3.883914 115.2670      68.09784      17.61857  32.21227 31.17363
  DiabetesPedigreeFunction      Age   Outcome
1                0.5972485 33.70303 0.5212121
2                0.4375705 33.11443 0.3018242
```

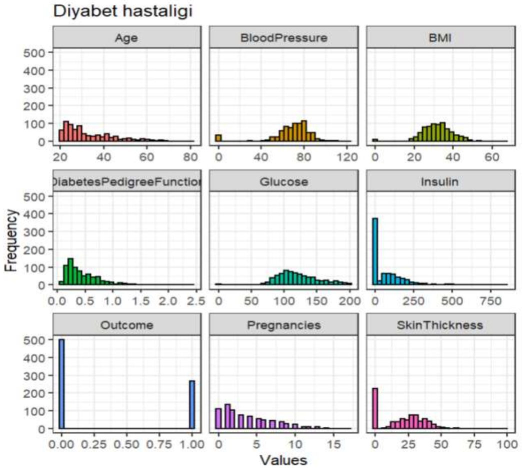
```
#diabetes adlı veri setinde Age ve BMI arasında bulunan ilişki grafikleştirildi.
plot(Age~BMI,diabetes)
with(diabetes,text(Age~Insulin,labels=BMI))
```



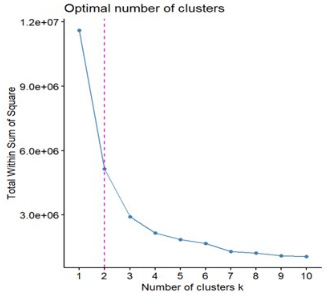
```
#clusplot grafiği oluşturuldu.
clusplot(diabetes, kmeans.küme$cluster, color=T, shade=T, labels=0, lines=0)
```



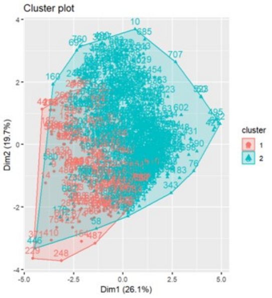
```
# Özelliklerin histogram grafikleri oluşturuldu.
diabetes %>%
  gather(Attributes, value, 1:9) %>%
  ggplot(aes(x=value, fill=Attributes)) +
  geom_histogram(colour="black", show.legend=FALSE) +
  facet_wrap(~Attributes, scales="free_x") +
  labs(x="Values", y="Frequency",
       title="Diyabet hastaligi") +
  theme_bw()
```



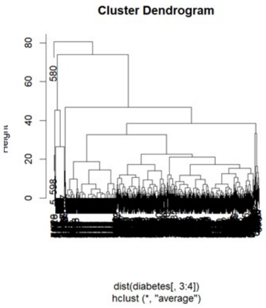
```
 # k=2 kırılma görüldüğünden grafik olarak gösterildi.

 fviz_nbclust(diabetes, kmeans, method="wss")+
   geom_vline(xintercept=2, linetype=2,col=6) #col=6 --> kırmızı renk
```



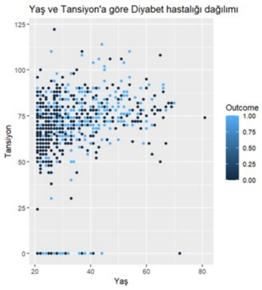
```
###### Kümeleme grafiği çizdirildi.
fviz_cluster(kmeans.küme, data=diabetes)
```



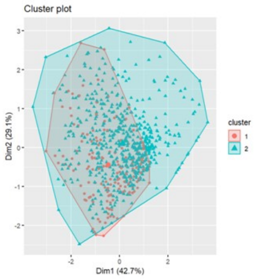
```
 #veri bölünerek dendrogram oluşturuldu.
 clusters <- hclust(dist(diabetes[, 3:4]), method = 'average')
 clusterCut <- cutree(clusters, 2)
 table(clusterCut, diabetes$Outcome)
 plot(clusters)
```

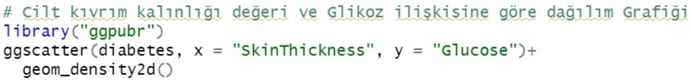


```
# Dağılım grafiği oluşturuldu.
dagilim <- ggplot(diabetes) +
  geom_point(aes(x = Age, y = BloodPressure, color =
                 Outcome)) +
  xlab("Yaş") +
  ylab("Tansiyon") +
  ggtitle("Yaş ve Tansiyon'a göre Diyabet hastalığı dağılımı")
print(dagilim)
```

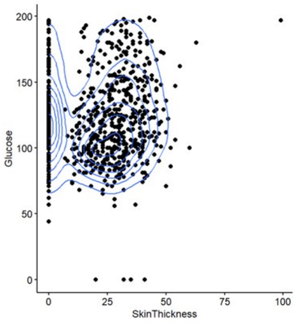


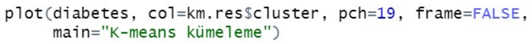
```
#Kümeleme grafiği çizdirildi.

diabetes.scaled<-scale(diabetes[,1:3])
fviz_cluster(kmeans.küme ,diabetes.scaled, geom = "point")
View(diabetes)
```



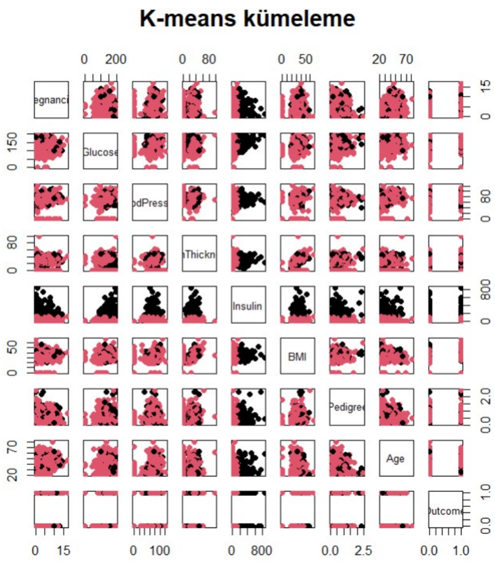


```
# Cilt kıvrım kalınlığı değeri ve Glikoz ilişkisine göre dağılım Grafiği
library("ggpubr")
ggscatter(diabetes, x = "SkinThickness", y = "Glucose")+
  geom_density2d()
```





```
plot(diabetes, col=km.res$cluster, pch=19, frame=FALSE,
     main="K-means kümeleme")
```



**Veri seti:**

**https://www.kaggle.com/datasets/saurabh00007/diabetescsv?select=diabetes.csv**

SENA BİLGİCİ