

BİL524 Veri Madenciliği

Dr. Öğr. Üy. İSMAİL YENİLMEZ

Gönül KILINÇ

**MAYIS 2024**

**ESKİŞEHİR**

#veri setini aktarma

canvas\_reklam <- read\_sav("C:/Users/CASPER/Downloads/canvas\_reklam.sav")

# Veri setini olusturma

n <- 200 # Gözlem sayisi

rahatsiz\_edici <- rnorm(n, mean = 10, sd = 2)

eglendirici <- rnorm(n, mean = 5, sd = 1)

bilgilendirici <- rnorm(n, mean = 3, sd = 0.5)

satin\_alma\_niyeti<- rnorm(n, mean = 7, sd = 1.5)

internet\_gunluk\_sure <- rnorm(n, mean = 15, sd = 3)

Katilimci\_no<- 3 + 2\*rahatsiz\_edici - 1.5\*eglendirici + 0.5\*bilgilendirici +

1.8\*satin\_alma\_niyeti - 1.2\*internet\_gunluk\_sure + rnorm(n, mean = 0, sd = 2)

# Olusturulan veri setini bir veri çerçevesine dönüştürme

canvas\_reklam <- data.frame(Katilimci\_no, rahatsiz\_edici, eglendirici,

bilgilendirici, satin\_alma\_niyeti, internet\_gunluk\_sure)

#veriyi görmek için view fonksiyonu kullanma

View(canvas\_reklam)

#ilk altı satiri görmek için head fonksiyonu kullanma

head(canvas\_reklam)

# rahatsiz\_edici degişkenini faktörel veri türüne dönüştürme

canvas\_reklam$rahatsiz\_edici<- as.factor(canvas\_reklam$rahatsiz\_edici)

#print fonksiyonu ile sadece rahatsiz edici degiskenini görme

print(rahatsiz\_edici)

#dplyr ile veri manipülasyonu

library(dplyr)

#select komutu ile rahatsiz edici degiskeninde ilk on veriyi görme

select(canvas\_reklam, rahatsiz\_edici)

#names fonksiyonu ile canvas reklam veri setine ait degisken isimlerini görme

names(canvas\_reklam)

#Ayni islemi colnames fonksiyonu ile yapma

colnames(canvas\_reklam)

# rahatsiz edici değgiskenkenini tam sayiya dönüştürme

canvas\_reklam$rahatsiz\_edici=as.integer(canvas\_reklam$rahatsiz\_edici)

print(rahatsiz\_edici)

#rahatsiz edici degiskenini numeric yapma

canvas\_reklam$rahatsiz\_edici <- as.numeric(canvas\_reklam$rahatsiz\_edici)

#str ile veri yapilarini kontrol etme

str(canvas\_reklam)

#mice paketi ile eksik veri tamamlama

#install.packages("mice")

install.packages("mice")

library(mice)

input\_data=nhanes

summary(input\_data)

#as.factor() fonksiyonu ile hyp degiskeninin veri türünü kategorik yapma

input\_data$hyp=as.factor(input\_data$hyp)

# eksik verileri tespit etmek için is.na() fonksiyonu kullanma

is.na(input\_data$hyp)

#eksik degerleri ortalama ile tamamlama

#which fonksiyonu TRUE olan yani bos olan degerleri doldurma.

#mean fonksiyonu ortalamayi temsil eder

#na.rm = TRUE bos degerlerin göz ardi edilmesi için kullanilir.

#na.rm=FALSE ise bos degerlerin dikkate alinması için kullanilir.

input\_data$bmi[which(is.na(input\_data$bmi))]=mean(input\_data$bmi, na.rm = TRUE)

#eksik veri tamamlama islemini chl degiskeni kullanilarak medyan ile tamamlama

input\_data$chl[which(is.na(input\_data$chl))]=median(input\_data$chl, na.rm = TRUE)

#mice paketi ile eksik verileri tamamlama

# yedek veri olusturma

input\_dt2=nhanes

# input\_data ve input\_ dt2 ayni oldugunu kontrol etmek "==" fonksiyonu kullanma

input\_data==input\_dt2

#hyp degikenini faktvr veri türü yapma

input\_dt2$hyp=as.factor(input\_dt2$hyp)

#methods(mice)= "mice" adli fonksiyon için mevcut olan yöntemleri

#veya metodlari listeleyen fonksiyonlari görme

methods(mice)

#"input\_dt2" veri setindeki eksik verileri doldurmak için MICE

#(Multiple Imputation by Chained Equations) yöntemini kullanarak bir imputasyon

#islemi gerçeklestirme.

my\_imp= mice(input\_dt2, m=5, method = c("","pmm","logreg","pmm"),maxit = 20)

# eksik verilerin oldugu veri seti"input\_dt2"

#m=5 parametresi, bes kez (m=5) eksik veri tamamlanacagini belirtir

#method = c("","pmm","logreg","pmm") parametresi, her bir s|tun igin

#kullanılacak olan veri tamamlama yö ntemlerini belirtir.

#Bos bir dize ("") kullanilarak, veri setinde eksik veri olmayan sütunlar belirtilir.

#maxit = 20 parametresi, her bir iterasyonda maksimum

#20 iterasyon yapilacagini belirtir.

summary(input\_dt2$bmi)

#input\_2 verisetinin vzetini gvsterir.

my\_imp$imp$bmi

#"my\_imp" adli MICE yöntemi ile yapilan eksik veri tamamlama islemi sonucunda

#elde edilen sonuçlar için "bmi" degiskenine ait tahmin edilen degerlerin

#bulundugu matrisi ifade eder.

final\_clean\_ds=complete(my\_imp,5)

#bu islem ile eksik verilerin tamamlanan veri seti ortaya çikar.

#yeni isleme baslamak için önce rm() ile daha önceki environmenti temizleme

rm(list=ls())

# Hazir veri çekme ve okuma

#excel verisini okuma için "readr" paketini indirme

#install.packages("readr")

library(readr)

#help sayfasinin ekrana gelmesi için

library(help=readr)

# "package ?" paket ismi ile de help sayfasında görüntüleme

package ? readr

#indirilenlerde olan "kaggle" dosyasini okuma

kaggle <- read\_csv("C:/Users/CASPER/Downloads/kaggle.zip")

#head fonksiyonu ile ilk alti satiri görüntüleme

head(kaggle)

# Kutu grafigi olusturma

veri=rnorm(100, mean = 1,sd=2)

boxplot(veri)

#sort küçükten büyüğe siralama

canvas\_reklam<- sort(rahatsiz\_edici)

print(canvas\_reklam)

# Z-puanı hesaplama

z\_scores <- scale(canvas\_reklam)

#z skorlarini görme.

View(z\_scores)

# Aykiri degerleri belirleme

install.packages("outliers")

library(outliers)

identify\_outliers(canvas\_reklam)

# örnek veri seti olusturma

data <- rnorm(1000, mean = 0, sd = 1)

hist(data)

#histogram ile normal dagilim görmek için hist fonksiyonu kullanma

hist(canvas\_reklam)

# kutu grafigi olusturma

boxplot(canvas\_reklam)

#Örnek veri olusturma

veri=rnorm(100, mean = 1,sd=2)

#kutu grafigi olusturma

boxplot(veri)

# Q-Q plot olusturma

qqnorm(veri)#normal dagilim için kullanilir

qqline(veri)# normal dagilim üzerine çizgi çizmek

qqnorm(canvas\_reklam)

qqline(canvas\_reklam)

#ortalam, medyan, maksimum ve minimum degerleri gömek için summary kullanma

summary(veri)

set.seed(123)

x <- rnorm(1000) # Bagimsiz degisken

y <- 2\*x + rnorm(1000, mean = 1, sd = 2) # Bagimli degisken

# Korelasyon katsayisini hesaplama

correlation <- cor(x, y)

#ekranda görmek için view fonksiyonunu kullanma

View(correlation)

#consolda görmek için print fonksiyonu kullanma

print(correlation)

#multicollinearity

rahatsiz\_edici <- rnorm(100)

bilgilendirici <- rnorm(100)

satin\_alma\_niyeti <- rnorm(100)

#istatistiksel model olusturma

model<-lm(Katilimci\_no~rahatsiz\_edici+bilgilendirici+satin\_alma\_niyeti)

#model özetini görme

summary(model)

#veri setindeki degiskenler arasindaki iliskiyi görmek corrlation\_matrix komutu kullanma

correlation\_matrix <- cor(data.frame(rahatsiz\_edici, bilgilendirici, satin\_alma\_niyeti))

#Korelasyonu görmek için

print(correlation\_matrix)

# corrplot paketini yükleyerek korelasyon matrisini görsellestirmek

install.packages("corrplot")

library(corrplot)

# Korelasyon matrisini görsellestirme

#correlation\_matrix: görsellestirmek istediginiz korelasyon matrisi.

#method = "color": Bu, korelasyon matrisinin nasil görsellestirilecegini

#belirten bir parametredir

corrplot(correlation\_matrix, method = "color")

# Bagimsiz degiskenlerin grafigini gizme

#par(mfrow=c(1,3)): Bu kod, grafiklerin düzenini ayarlar.

#mfrow parametresi, çizilecek grafiklerin sayisini ve düzenini belirler.

#1=satir, 3=sütunu gösteririr.

#plot dagilma grafigini ifade eder.

#main, xlab, ve ylab baslik, x-ekseni etiketi ve y ekseni etiketi

#için metin saglar.

#col = "blue" ve pch ise grafiğin rengini ve nokta tipini belirtir.

par(mfrow=c(1,3))

plot(rahatsiz\_edici, Katilimci\_no, main = "rahatsiz\_edici vs.Katilimci\_no", xlab = "rahatsiz\_edici", ylab = "Katilimci\_no", col = "blue", pch = 16)

plot(bilgilendirici, Katilimci\_no, main = "bilgilendirici vs. Katilimci\_no", xlab = "bilgilendirici", ylab = "Katilimci\_no", col = "red", pch = 16)

plot(satin\_alma\_niyeti, Katilimci\_no, main = "satin\_alma\_niyeti vs. Katilimci\_no", xlab = "satin\_alma\_niyeti", ylab = "Katilimci\_no", col = "green", pch = 16)

#veri seti olusturma

rahatsiz\_edici <- rnorm(100)

bilgilendirici <- rnorm(100)

satin\_alma\_niyeti <- rnorm(100)

Katilimci\_no <- 2\*rahatsiz\_edici + 3\*bilgilendirici - 1.5\*satin\_alma\_niyeti + rnorm(100, mean = 0, sd = 0.5)

# Veriyi standartlastirarak ortalama "0" deger almasi ve standart sapma"1" olmasi

rahatsiz\_edici\_standardized <- scale(x1)

bilgilendirici\_standardized <- scale(x2)

satin\_alma\_niyeti\_standardized <- scale(x3)

Katilimci\_no\_standardized <- scale(y)

#summary fonksiyonu ile standartlastirilmis degiskenleri görme

summary(rahatsiz\_edici\_standardized)

summary(bilgilendirici\_standardized)

summary(satin\_alma\_niyeti\_standardized)

summary(Katilimci\_no\_standardized)

# caTools paketini yükleyerek siniflandirma ve regresyon için veri ön

#islemi yapma

install.packages("caTools")

library(caTools)

#Örnek veri seti olusturma

rahatsiz\_edici <- rnorm(100)

bilgilendirici <- rnorm(100)

satin\_alma\_niyeti <- rnorm(100)

Katilimci\_no <- 2\*rahatsiz\_edici + 3\*bilgilendirici - 1.5\*satin\_alma\_niyeti + rnorm(100, mean = 0, sd = 0.5)

# Veriyi test ve egitim gibi alt kümelerine bvlme islemi yapma

split <- sample.split(Katilimci\_no, SplitRatio = 0.7)

train\_data <- subset(data.frame(rahatsiz\_edici, bilgilendirici, satin\_alma\_niyeti, Katilimci\_no), split == TRUE)

test\_data <- subset(data.frame(rahatsiz\_edici, bilgilendirici, satin\_alma\_niyeti, Katilimci\_no), split == FALSE)

#verinin boyutlarını görmek için dim fonksiyonu kullanma

dim(train\_data)

dim(test\_data)