

SAP Projekt

Statistički Analizirani

2025-12-10

Motivacija i opis problema

Moždani udar je hitno medicinsko stanje koje nastaje nakon poremećaja cirkulacije u mozgu. U Hrvatskoj je drugi najveći uzrok smrti te prvi najveći uzrok invaliditeta. Rizik od moždanog udara kod pojedine osobe može ovisiti o više faktora pa se u medicini koriste razni modeli za njegovu procjenu. Nije strana ni uporaba alata zasnovanih na umjernoj inteligenciji. Predikcija moždanog udara bitna je za pravovremenu identifikaciju rizika koje omogućava pravovremene mjere prevencije i osvještavanje populacije.

Opis skupa podataka

Prikupljeni skup podataka sadrži kliničke podatke o pacijentima i informacije o moždanom udaru. Za svakog pacijenta navedena je vrijednost 12 značajki grupiranih u 4 kategorije: demografski podaci, zdravstveni podaci, fiziološki podaci i životne navike. Cilj je uočiti povezanost izmjerenih podataka i rizika od moždanog udara. Skup podataka sadržava 5,110 zapisa o pacijentima od kojih je njih 249 doživjelo moždani udar. Udio pacijenata koji su doživjeli moždani udar iznosi 4.87%

- id: jedinstveni identifikator pacijenta
- gender: spol pacijenta (Male, Female)
- age: dob pacijenta
- hypertension: oznaka koja daje informaciju o tome ima li pacijent visoki tlak (0, 1)
- heart_disease: oznaka koja daje informaciju ima li pacijent neku srčanu bolest (0, 1)
- ever_married: odgovara na pitanje je li pacijent ikad bio u braku (No, Yes) work_type: tip zaposlenja (children, Govt_job, Never_worked, Private, Self-employed)
- Residence_type: tip prebivališta u kojem živi pacijent (Rural, Urban)
- avg_glucose_lvl: prosječna razina glukoze u krvi (mg/dL)
- bmi: indeks tjelesne mase koji predstavlja odnos visine i težine pacijenta
- smoking_status: opis pacijentovog odnosa s pušenjem cigareta (formerly smoked, never smoked, smokes, Unknown)
- stroke: oznaka koja daje informaciju je li pacijent doživio moždani udar (0, 1)

Učitavanje i pregled podataka

```
data <- read.csv("data.csv")
head(data)
```

```
##      id gender age hypertension heart_disease ever_married  work_type
## 1  9046  Male  67             0              1           Yes   Private
## 2 51676 Female  61             0              0           Yes Self-employed
## 3 31112  Male  80             0              1           Yes   Private
```

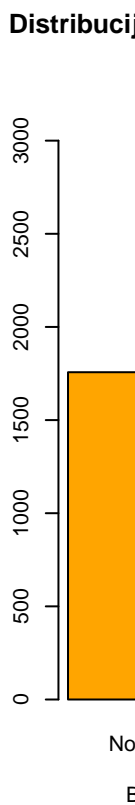
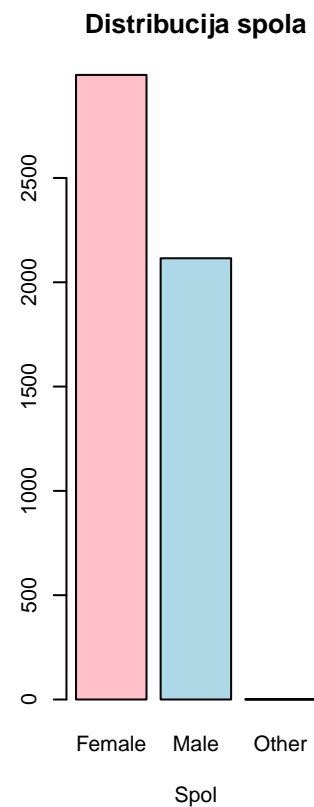
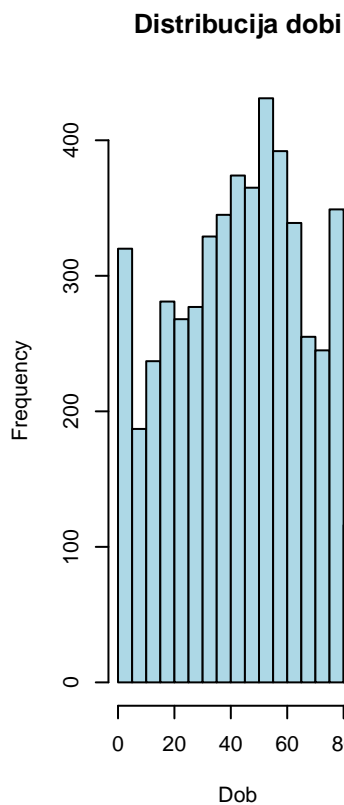
```
## 4 60182 Female 49          0          0      Yes      Private
## 5 1665 Female 79          1          0      Yes Self-employed
## 6 56669 Male 81          0          0      Yes      Private
##   Residence_type avg_glucose_level  bmi  smoking_status  stroke
## 1      Urban      228.69 36.6  formerly smoked      1
## 2      Rural      202.21 N/A   never smoked      1
## 3      Rural      105.92 32.5  never smoked      1
## 4      Urban      171.23 34.4      smokes      1
## 5      Rural      174.12 24   never smoked      1
## 6      Urban      186.21 29  formerly smoked      1
```

```
str(data)
```

```
## 'data.frame': 5110 obs. of 12 variables:
## $ id : int 9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
## $ gender : chr "Male" "Female" "Male" "Female" ...
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...
## $ hypertension : int 0 0 0 0 1 0 1 0 0 0 ...
## $ heart_disease : int 1 0 1 0 0 0 1 0 0 0 ...
## $ ever_married : chr "Yes" "Yes" "Yes" "Yes" ...
## $ work_type : chr "Private" "Self-employed" "Private" "Private" ...
## $ Residence_type : chr "Urban" "Rural" "Rural" "Urban" ...
## $ avg_glucose_level: num 229 202 106 171 174 ...
## $ bmi : chr "36.6" "N/A" "32.5" "34.4" ...
## $ smoking_status : chr "formerly smoked" "never smoked" "never smoked" "smokes" ...
## $ stroke : int 1 1 1 1 1 1 1 1 1 1 ...
```

Iz generiranih prikaza moguće je vidjeti neke od vrijednosti koje poprimaju pojedini atributi vezani uz pacijenta, ali i tip podataka koji prikazuju. Primjetljivo je kako se atribut bmi vodi kao niz znakova iako semantički predstavlja decimalni broj. Pri izradi dijagrama i u budućim računima potrebno je pripaziti da se bmi ne prikaže kao kategorični atribut.

```
par(mfrow = c(1,3))
hist(data$age , main = "Distribucija dobi",
     xlab = "Dob",
     col = "lightblue", border = "black")
barplot(table(data$gender),
      main = "Distribucija spola", xlab = "Spol",
      col = c("pink","lightblue","black"),border="black")
barplot(table(data$ever_married),
      main = "Distribucija bralnog statusa",
      xlab = "Bralni status",
      col = c("orange","lightblue"),border="black")
```



Prikaz demografskih podataka

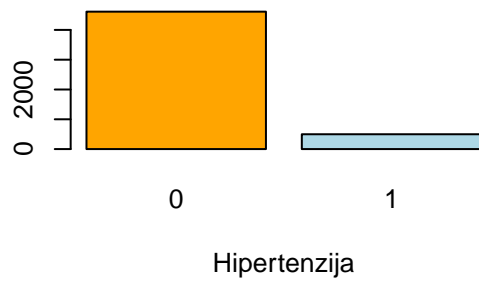
Prikaz zdravstvenih podataka Kao što je prethodno navedeno, potrebno je pripremiti atribut bmi.

```
bmi<-data$bmi
```

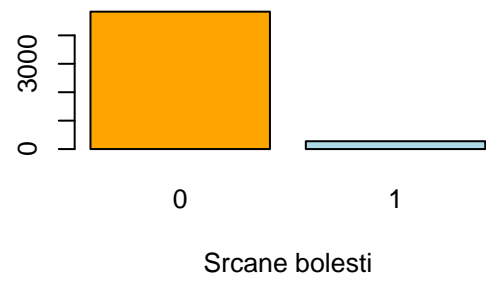
```
bmi[bmi == "N/A"] <- NA
bmi <- as.numeric(bmi)
```

```
par(mfrow = c(2,2))
barplot(table(data$hypertension),
        main = "Distribucija hipertenzije",
        xlab = "Hipertenzija",
        col = c("orange","lightblue"),border="black")
barplot(table(data$heart_disease),
        main = "Distribucija srcanih bolesti",
        xlab = "Srcane bolesti",
        col = c("orange","lightblue"),border="black")
hist(data$avg_glucose_level ,
     main = "Distribucija kolicine glukoze u krvi",
     xlab = "Kolicina glukoze u krvi [mg/dL]",
     col = "lightblue", border = "black")
hist(bmi , main = "Distribucija indeksa BMI",
     xlab = "BMI", col = "lightblue",
     border = "black")
```

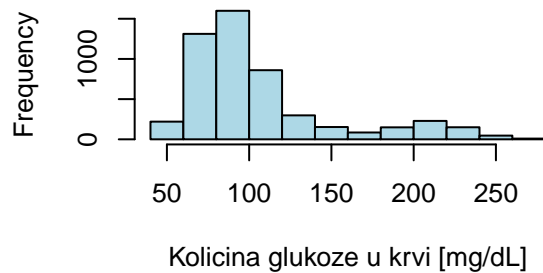
Distribucija hipertenzije



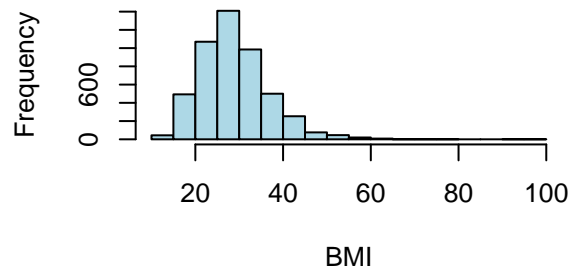
Distribucija srčanih bolesti



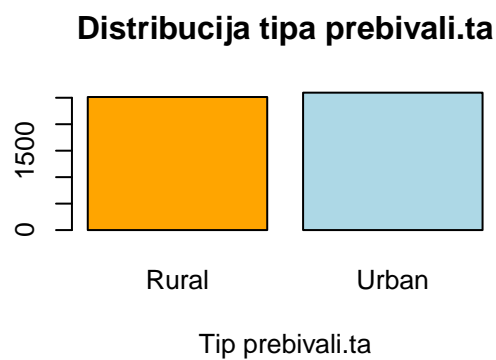
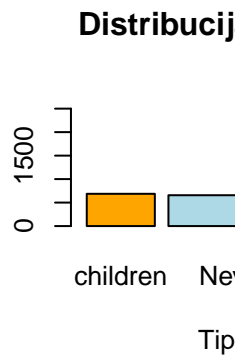
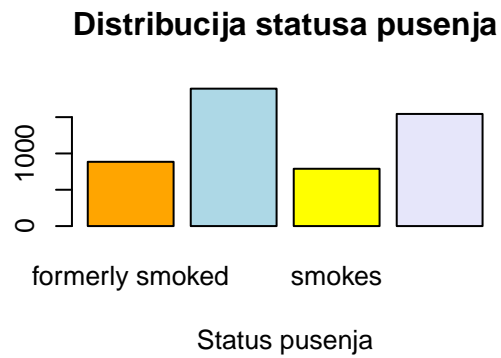
Distribucija kolicine glukoze u krvi



Distribucija indeksa BMI



```
par(mfrow = c(2,2))
barplot(table(data$smoking_status),
        main = "Distribucija statusa pusenja",
        xlab = "Status pusenja",
        col = c("orange","lightblue","yellow","lavender"),border="black")
barplot(table(data$work_type),
        main = "Distribucija tipa zaposlenja",
        xlab = "Tip zaposlenja",
        col = c("orange","lightblue","yellow","lavender","aquamarine"),border="black")
barplot(table(data$Residence_type),
        main = "Distribucija tipa prebivališta",
        xlab = "Tip prebivališta",
        col = c("orange","lightblue","yellow","lavender"),border="black")
```



Prikaz podataka o životnim navikama

Postoji li statistički značajna razlika u prosječnoj razini glukoze između pacijenata sa i bez moždanog udara?

```
data_stroke <- data[data$stroke == 1, ]$avg_glucose_level
data_no_stroke <- data[data$stroke == 0,]$avg_glucose_level
summary(data_stroke)
```

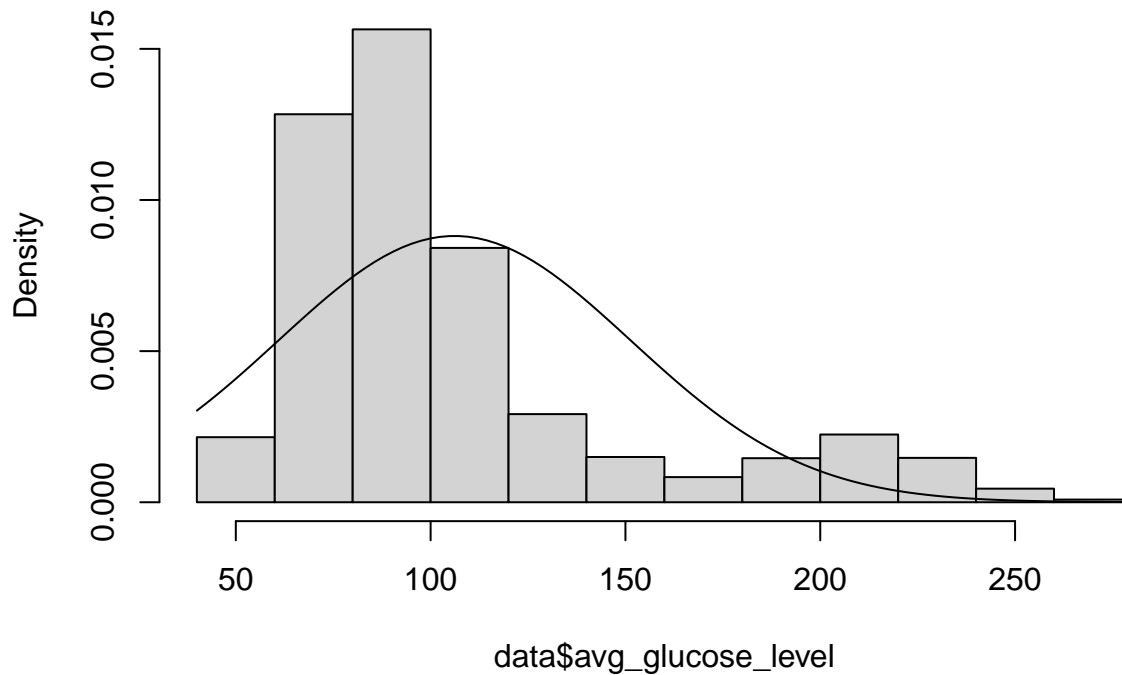
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  56.11   79.79   105.22   132.54   196.71   271.74
```

```
summary(data_no_stroke)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  55.12   77.12   91.47   104.80   112.83   267.76
```

```
hist(data$avg_glucose_level, freq = FALSE)
curve(dnorm(x, mean(data$avg_glucose_level), sd(data$avg_glucose_level)), add = TRUE)
```

Histogram of data\$avg_glucose_level



```
set.seed(67)

B <- 10000

boot_diff <- numeric(B)

for (i in 1:B) {
  boot_data_stroke <- sample(data_stroke, replace = TRUE)
  boot_data_no_stroke <- sample(data_no_stroke, replace = TRUE)
  boot_diff[i] <- mean(boot_data_stroke) - mean(boot_data_no_stroke)
}

# Observed mean difference
obs_diff <- mean(data_stroke) - mean(data_no_stroke)

# 95% bootstrap confidence interval
ci <- quantile(boot_diff, c(0.025, 0.975))

list(
  observed_mean_difference = obs_diff,
  ci_95 = ci
)
```

```
## $observed_mean_difference
## [1] 27.74923
##
```

```
## $ci_95
##      2.5%      97.5%
## 19.94809 35.74191
```

Postoji li povezanost između statusa pušenja i nastanka moždanog udara?

Da bi vidjeli povezanost statusa pušenja i nastanka moždanog udara trebamo primijeniti test hi-kvadrat povezanosti.

Za početak trebamo maknuti unknown s tablice jer neznamo njihov smoking status te onda primjenjujemo test.

```
data$smoking_status <- as.factor(data$smoking_status)
data$stroke <- as.factor(data$stroke)

tablica <- table(data$smoking_status, data$stroke)
tablica <- tablica[-4, ] # izbacujemo unknown pušače
tablica
```

```
##
##              0      1
## formerly smoked 815   70
## never smoked   1802  90
## smokes         747   42
```

```
chi_rezultat <- chisq.test(tablica)
chi_rezultat
```

```
##
## Pearson's Chi-squared test
##
## data:  tablica
## X-squared = 11.436, df = 2, p-value = 0.003285
```

```
chi_rezultat$expected
```

```
##
##              0      1
## formerly smoked 834.8682 50.13180
## never smoked   1784.8256 107.17443
## smokes         744.3062  44.69377
```

Vidimo da je $p\text{-value} = 0.003285$ te iz njega možemo izbaciti prvu hipotezu tojest vidimo da je smoking_status povezan s moždanim udarom.

Za sljedeći primjer ćemo izmijeniti tablicu koja zadržava smoking_status i stroke. Ljude koje su prestali pušiti i one koji puše ćemo staviti u jedan red skupa, a ostale tojest one koji nisu nikada pušili ćemo staviti u drugi red.

```
nova_tablica2 <- table(data$smoking_status, data$stroke)

nova_tablica2 <- nova_tablica2[-4, ]
nova_tablica2
```

```
##
##           0    1
##  formerly smoked 815  70
##  never smoked   1802  90
##  smokes         747  42
```

```
nova_tablica2[2,] <- nova_tablica2[2, ] + nova_tablica2[3, ]
nova_tablica2 <- nova_tablica2[-3, ]
```

```
rownames(nova_tablica2)[1] <- "never smoked"
rownames(nova_tablica2)[2] <- "was smoking/smokes"
```

```
nova_tablica2
```

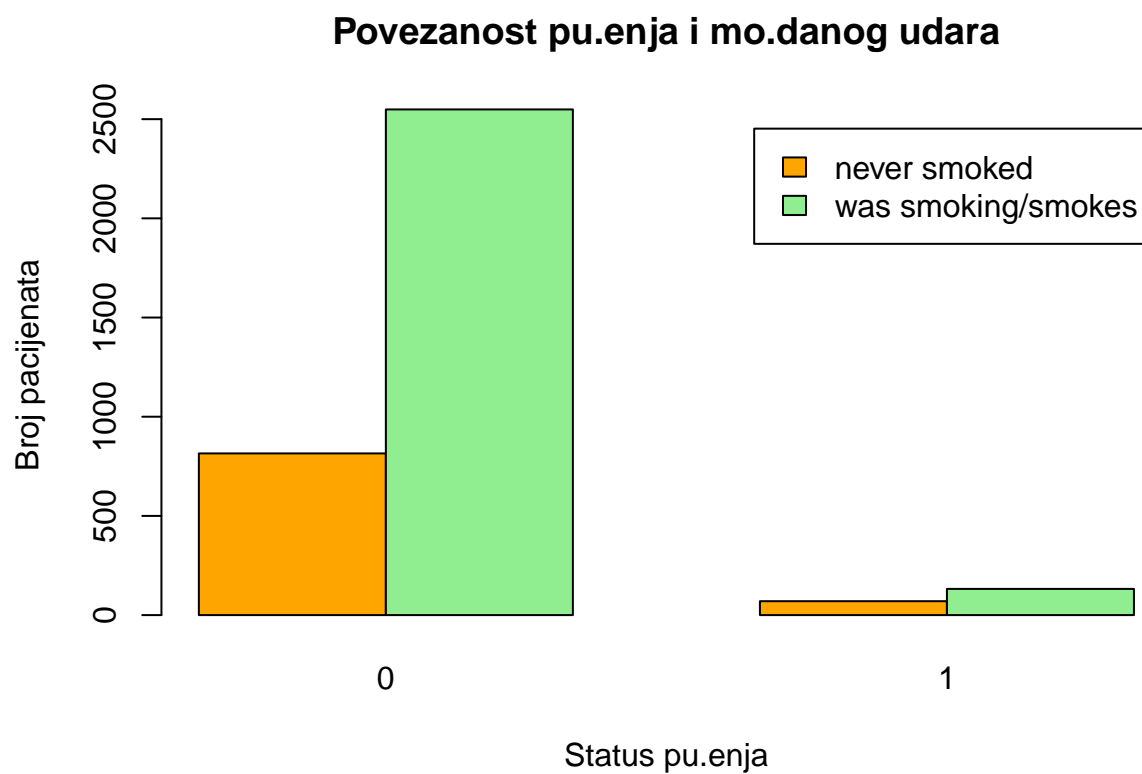
```
##
##           0    1
##  never smoked      815  70
##  was smoking/smokes 2549 132
```

```
chi_rezultat2 <- chisq.test(nova_tablica2)
```

```
chi_rezultat2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  nova_tablica2
## X-squared = 10.551, df = 1, p-value = 0.001162
```

```
barplot(nova_tablica2,
beside = TRUE,
legend = TRUE,
col = c("orange", "lightgreen"),
main = "Povezanost pušenja i moždanog udara",
xlab = "Status pušenja",
ylab = "Broj pacijenata")
```

Vidimo da je $p\text{-value} = 0.001162$ iz čega možemo zaključiti da i iz ovog testa postoji povezanost između statusa pušenja i moždanog udara.