Predikcija moždanog udara

13. prosinca 2025.

1 Upoznavanje s podatcima

U ovom se odjeljku učitavaju dani podatci te se provodi njihovo čišćenje i priprema za daljnju analizu.

1.1 Učitavanje podataka

U sljedećem koraku učitavamo skup podataka o moždanom udaru iz CSV datoteke te provodimo osnovnu analizu o podudarnosti tipova atributa.

```
# Učitavanje danog skupa podataka
podatci <- read.csv("healthcare-dataset-stroke-data.csv")
```

Nakon uspješnog učitavanja, napravit ćemo kratak sažetak svih atributa učitanih podataka.

```
str(podatci)
```

```
## '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 ...
```

Primijetimo kako svi atributi, izuzev **bmi** atributa, imaju odgovarajuće tipove podataka definirane u opisu podatkovnog skupa što je važno za daljnje provođenje postupka čišćenja

```
summary(podatci)
```

```
##        id           gender               age         hypertension
##  Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Mode  :character   Median :45.00   Median :0.00000
##  Mean   :36518                      Mean   :43.23   Mean   :0.09746
##  3rd Qu.:54682                      3rd Qu.:61.00   3rd Qu.:0.00000
##  Max.   :72940                      Max.   :82.00   Max.   :1.00000
##  heart_disease     ever_married        work_type         Residence_type
##  Min.   :0.00000   Length:5110        Length:5110        Length:5110
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
```

```
##  Mean   :0.05401
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  avg_glucose_level     bmi            smoking_status         stroke
##  Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000
##  1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000
##  Median : 91.89    Mode  :character   Mode  :character   Median :0.00000
##  Mean   :106.15                                          Mean   :0.04873
##  3rd Qu.:114.09                                          3rd Qu.:0.00000
##  Max.   :271.74                                          Max.   :1.00000
```

1.2 Čišćenje podataka

Ovaj se odjeljak posvećuje svakome podatkovnom atributu kako bi se utvrdila valjanost podataka.

1.2.1 Varijabla `id`

```
sum(duplicated(podatci$id))
```

```
## [1] 0
```

Provjera postije li podatci pod istim ID-om. Svi su podatci označeni jedinstvenim identifikatorom te u ovome koraku nema potrebe za bilo kakvom manipulacijom nad njima.

1.2.2 Varijabla `gender` (spol)

```
unique(podatci$gender)
```

```
## [1] "Male"   "Female" "Other"
```

```
sum(is.na(podatci$gender))
```

```
## [1] 0
```

```
table(podatci$gender)
```

```
##
## Female   Male  Other
##   2994   2115      1
```

```
cat("Postotak \'Other\' kategorije: ", round(sum(podatci$gender == "Other") * 100 / length(podatci$gende
```

```
## Postotak 'Other' kategorije:  0.02 %
```

```
podatci <- podatci[podatci$gender != "Other",]
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti. S obziorm na to da je zastupljenost podataka za kategoriju "Others" premalena (≈ 0.02, tj. frekvencija pojavljivanja je 1), teško ju je modelirati te je izbačen odgovarajući podatak.

1.2.3 Varijabla `age` (dob)

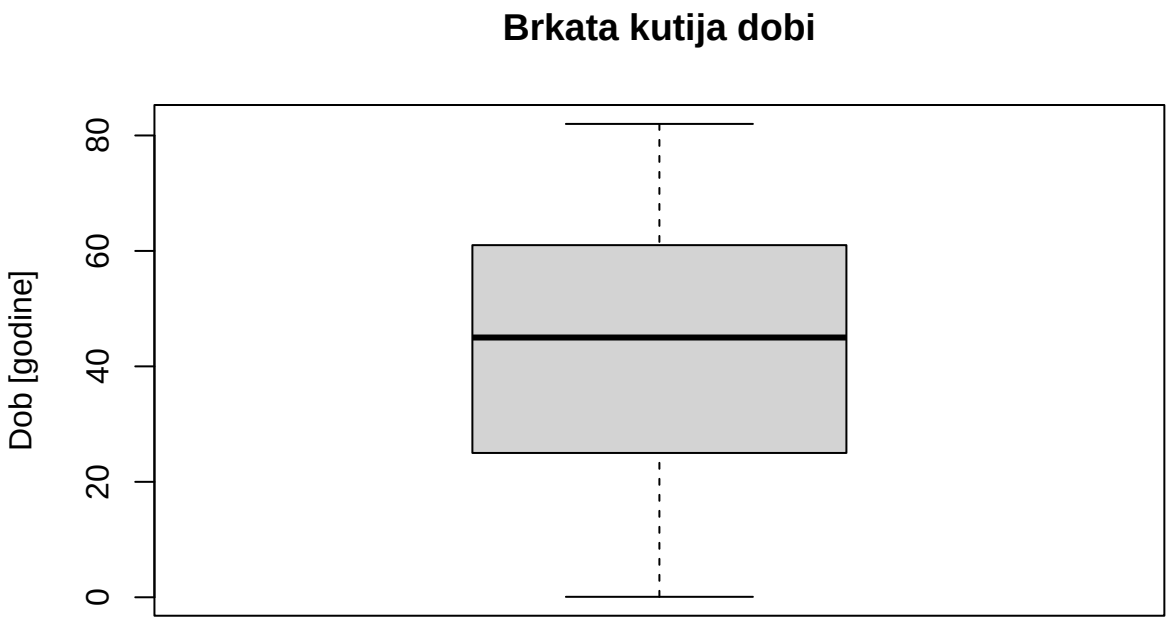
```
summary(podatci$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.08   25.00   45.00   43.23   61.00   82.00
```

```r
sum(is.na(podatci$age))
```

```
## [1] 0
```

```r
boxplot(podatci$age, main="Brkata kutija dobi", range=1.5, ylab="Dob [godine]")
```



Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti, stršeće vrijednosti nisu detektirane te se podatci u atributu *age* ne trebaju dalje čistiti.

1.2.4 Varijabla `hypertension` (visoki krvni tlak)

```r
unique(podatci$hypertension)
```

```
## [1] 0 1
```

```r
sum(is.na(podatci$hypertension))
```

```
## [1] 0
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti te se podatci u atributu *hypertension* ne trebaju dalje čistiti.

1.2.5 Varijabla `heart_disease` (srčane bolesti)

```r
unique(podatci$heart_disease)
```

```
## [1] 1 0
```

```
sum(is.na(podatci$heart_disease))
```

```
## [1] 0
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti te se podatci u atributu *heart_disease* ne trebaju dalje čistiti.

1.2.6 Varijabla `ever_married` (bračni status)

```
unique(podatci$ever_married)
```

```
## [1] "Yes" "No"
```

```
sum(is.na(podatci$ever_married))
```

```
## [1] 0
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti te se podatci u atributu *ever_married* ne trebaju dalje čistiti.

1.2.7 Varijabla `work_type` (tip zaposlenja)

```
unique(podatci$work_type)
```

```
## [1] "Private"       "Self-employed" "Govt_job"      "children"
## [5] "Never_worked"
```

```
sum(is.na(podatci$work_type))
```

```
## [1] 0
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti te se podatci u atributu *work_type* ne trebaju dalje čistiti.

1.2.8 Varijavla `residence_type` (tip prebivališta)

```
unique(podatci$Residence_type)
```

```
## [1] "Urban" "Rural"
```

```
sum(is.na(podatci$Residence_type))
```

```
## [1] 0
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti te se podatci u atributu *residence_type* ne trebaju dalje čistiti.

1.2.9 Varijabla `avg_glucose_level` (prosječna razina glukoze)

```
summary(podatci$avg_glucose_level)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.12   77.24   91.88  106.14  114.09  271.74
```

U specifikacij zadatka nisu dane restrikcije na domenu varijable *avg_glucose_level*, stoga numeričke vrijednosti u podatkovnom skupu nije moguće usporeditit ni s čim te će se samo uzeti kao valjane.

1.2.10 Varijabla bmi (indeks tjelesne mase)

Kao što je već otkriveno prilikom učitavanja podataka, potrebno je promijeniti tip varijable bmi. Kako bi navedeno bilo moguće, nedostajuće vrijednosti, koje su u podatcima obilježene znakovnim nizom "N/A", moramo pretvoriti u NA objekt poznat R-u. Zatim možemo provesti pretvorbu podataka.

```
# br nepostojećih vrijednosti
length(podatci$bmi[podatci$bmi == "N/A"])
```

```
## [1] 201
```

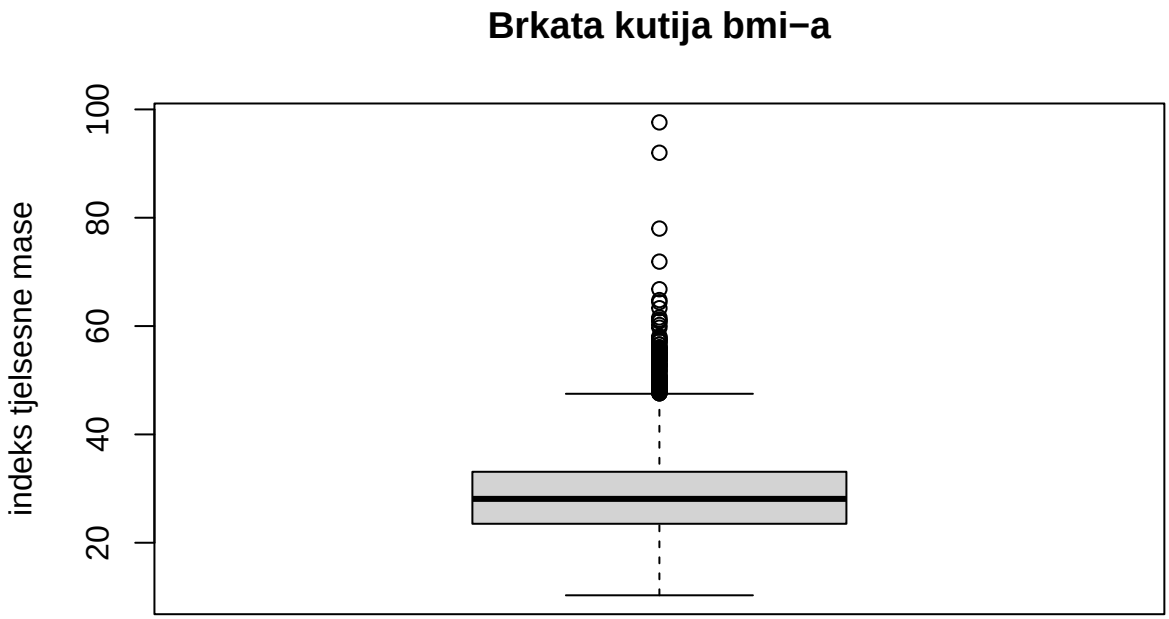
```
podatci$bmi[podatci$bmi == "N/A"] <- NA

podatci$bmi <- as.numeric(podatci$bmi)

summary(podatci$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.30   23.50   28.10   28.89   33.10   97.60     201
```

```
boxplot(podatci$bmi, main="Brkata kutija bmi-a", range=1.5, ylab="indeks tjelesne mase")
```




```
cat("Br. stršećih vrijednosti:", length(boxplot.stats(podatci$bmi, coef = 1.5)$out))
```

```
## Br. stršećih vrijednosti: 110
```

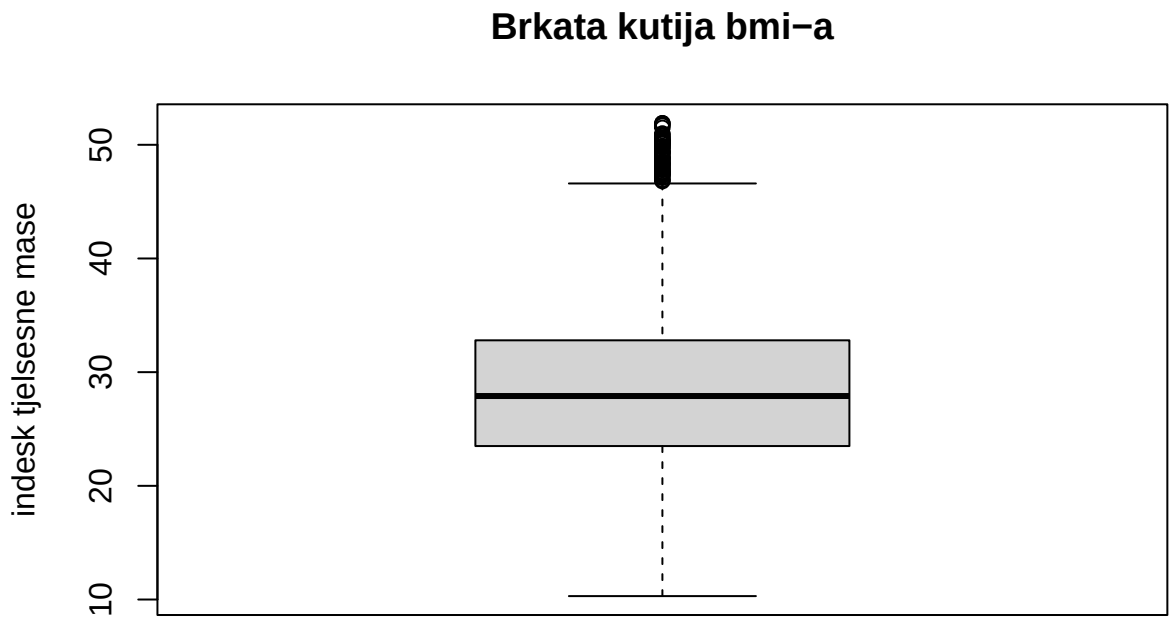
Nakon nužne zamjene znakovnog niza "N/A" u NA objekt i pretvorbe bmi podataka u numerički tip, može se izvršiti metoda `summary()` iz koje vidimo da se sve postojeće vrijednosti u podatkovnom skupu sada podudaraju s domenom atributa specificiranom u zadatku, točno je identificirana 201 nepoznata vrijednost te

se podatci u atributu *bmi* ne trebaju dalje čistiti. Pomoću metode `boxplot()` vizualizirana je brkata kutija podataka varijable *bmi*, utvrđeno je postojanje 110 stršećih vrijednosti koje su sve iznad gornjeg izdanka brkate kutije. Kako bi se smanjio utjecaj stršećih vrijednosti, odbačeni su svi podatci iz danog skupa za koje varijabla *bmi* poprima uglavnom nerealne vrijednosti (≥ 52).

```
podatci <- podatci[podatci$bmi < 52 | is.na(podatci$bmi),]
summary(podatci$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.30   23.50   27.90   28.54   32.80   51.90     201
```

```
boxplot(podatci$bmi, main="Brkata kutija bmi-a", range=1.5, ylab="indesk tjelsesne mase")
```




```
cat("Br. stršećih vrijednosti:", length(boxplot.stats(podatci$bmi, coef = 1.5)$out))
```

```
## Br. stršećih vrijednosti: 61
```

Nakon uklanjanja nerealnih vrijednosti, vidljiv je manji utjecaj stršećih vrijednosti na aritmetički sredinu *bmi* podataka. Kako ne bi došlo do daljnjeg gubitka podataka, s nedostajaćim će se podatcima pristupati tako da im se pridruži vrijednost medijana. Za ovaj je postupak odabran medijan umjesto srednje vrijednosti upravo zbog nedostatka utjecaja gornjih stršećih vrijednosti na njega.

```
podatci$bmi[is.na(podatci$bmi)] <- median(podatci$bmi, na.rm = TRUE)
```

1.2.11 Varijabla `smoking_status`(status pušenja)

```
unique(podatci$smoking_status)
```

```
## [1] "formerly smoked" "never smoked"    "smokes"          "Unknown"
```

```r
sum(is.na(podatci$smoking_status))
```

```
## [1] 0
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, nepoznate vrijednosti postoje jer su u samoj definiciji dopuštenih realizacija varijable `smoking_status` te su označeni kao "Unknown" radije nego objektom NA. Podatci u atributu *smoking_status* se ne trebaju dalje čistiti.

1.2.12 Varijabla `stroke` (moždani udar)

```r
unique(podatci$stroke)
```

```
## [1] 1 0
```

```r
sum(is.na(podatci$stroke))
```

```
## [1] 0
```

Sve se postojeće vrijednosti u podatkovnom skupu podudaraju s domenom atributa specificiranom u zadatku, ne postoje nedostajeće i nepoznate vrijednosti te se podatci u atributu *stroke* ne trebaju dalje čistiti.

2 Literatura

1. Pintar, Damir; Programirajmo u R-u; 2025-11-13; URL: https://ratnip.github.io/OSP_2025/; pristupljeno 2025-12-14