MULTIVARIJANTNA I DUBINSKA ANALIZA PODATAKA

ZDRAVSTVENI POKAZATELJI DIJABETESA

CDC DIABETES HEALTH INDICATORS

JOSHUA LEE FLETCHER, NOA MIDŽIĆ
MENTORICA: PROF. DR. SC. JASMINKA DOBŠA
FAKULTET ORGANIZACIJE I INFORMATIKE
INFORMACIJSKO I PROGRAMSKO INŽENJERSTVO 1.4

SADRŽAJ

PODACI O DATASETU

- Behavioral risk factor surveillance system (BRFSS) – godišnja anketa CDC-a preko telefona
- 70,692 pojedinaca odgovorilo (opservacija)
  - 21 varijabli (podaci su odgovori ili izračunati s obzirom na odgovore)
  - 50% nema dijabetes
- Diabetes_binary
  - 0 – nema dijabetes, 1 - ima preddijabetes ili dijabetes

Varijable koje ćemo analizirati

Prikupljeni podaci za različite rizične i druge faktore za dijabetes – kolesterol, fizičko i mentalno zdravlje, pokriće zdravstvenim osiguranjem, ekonomski i socijalni status…

Varijabla	Značenje	Vrsta varijable	Vrijednosti koje poprima
HighChol	Razina kolesterola	Kvalitativna, nominalna	0 (nizak), 1 (visok)
CholCheck	Pregledan kolesterol u zadnjih 5 godina?	Kvalitativna, nominalna	0 (jest), 1 (nije)
BMI_Group	Indeks tjelesne mase	Numerička, diskretna	12-98
HvyAlcoholConsump	Žene >= 7 pića tjedno Muškarci >= 14 pića tjedno	Kvalitativna, nominalna	0 (ne), 1 (da)
AnyHealthcare	Ima li zdravstveno osiguranje?	Kvalitativna, nominalna	0 (ne), 1 (da)
NoDocbcCost	U zadnjih godinu dana, nemogućnost odlaska doktoru zbog cijene?	Kvalitativna, nominalna	0 (ne), 1 (da)
Sex	Spol	Kvalitativna, nominalna	0 (žena), 1 (muškarac)

Tablica 1. Varijable koje ćemo analizirati

Varijable koje ćemo analizirati

Varijabla	Značenje	Vrsta varijable	Vrijednosti koje poprima
HighChol	Razina kolesterola	Kvalitativna, nominalna	0 (nizak), 1 (visok)
CholCheck	Pregledan kolesterol u zadnjih 5 godina?	Kvalitativna, nominalna	0 (jest), 1 (nije)
BMI_Group	Indeks tjelesne mase	Numerička, diskretna	12-98
HvyAlcoholConsump	Žene >= 7 pića tjedno Muškarci >= 14 pića tjedno	Kvalitativna, nominalna	0 (ne), 1 (da)
AnyHealthcare	Ima li zdravstveno osiguranje?	Kvalitativna, nominalna	0 (ne), 1 (da)
NoDocbcCost	U zadnjih godinu dana, nemogućnost odlaska doktoru zbog cijene?	Kvalitativna, nominalna	0 (ne), 1 (da)
Sex	Spol	Kvalitativna, nominalna	0 (žena), 1 (muškarac)

Tablica 1. Varijable koje ćemo analizirati

Varijable koje ćemo analizirati

Varijabla	Značenje	Vrsta varijable	Vrijednosti koje poprima
Age_Group	AGE5GYR skala dobi: 1 = 18-24, 2 = 25-29, 3 = 30-34, 4 = 35-39, 5 = 40-44, 6 = 45-49, 7 = 50-54, 8 = 55-59, 9 = 60-64, 10 = 65-69, 11 = 70-74, 12 = 75-79, 13 = 80+ godina	Kvalitativna, redoslijedna	1-13
Education_Group	EDUCA skala obrazovanja: 1 = samo vrtić, 2 = osnovna, 3 = nešto srednje škole, 4 = srednja škola, 5 = fakultet 1-3 godine, 6 = fakultet 4 godine ili više	Kvalitativna, redoslijedna	1-6
Income	INCOME2 skala zarade: 1 = manje od 10.000 dolara, 2 = 10.000-15.000 dolara, 3 = 15.000-20.000 dolara, 4 = 20.000-25.000, 5 = 25.000-35.000, 6 = 35.000-50.000, 7 = 50.000-75.000, 8 = 75.000 ili više dolara	Kvalitativna, redoslijedna	1-8

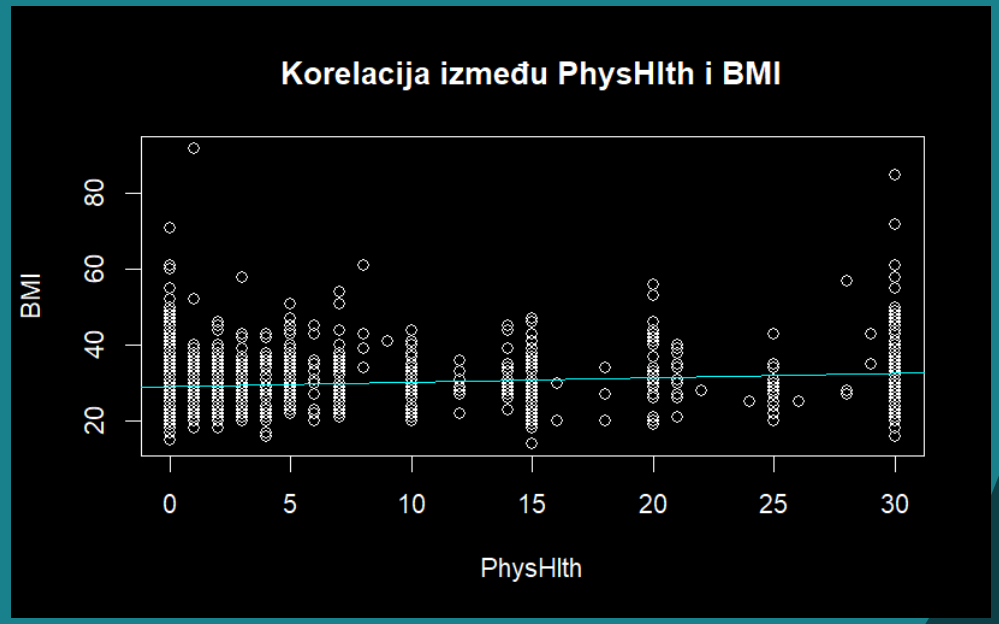
Tablica 1. Varijable koje ćemo analizirati

Varijable koje ćemo analizirati

Varijabla	Značenje	Vrsta varijable	Vrijednosti koje poprima
HighBP	Ima li povišen tlak krvi?	Kvalitativna, nominalna	0 (ne), 1 (da)
GenHlth	Osobna procjena općenitog zdravtsvenog stanja na skali 1-5: 1 = odlično, 2 = vrlo dobro, 3 = dobro, 4 = u redu, 5 = loše	Kvalitativna, redoslijedna	1-5

Tablica 1. Varijable koje ćemo analizirati

DESKRIPTIVNA STATISTIKA



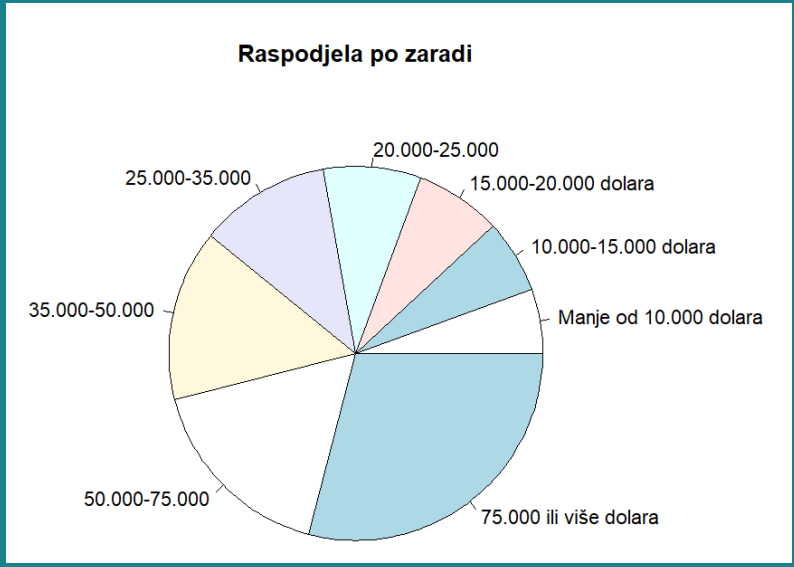
Slika 1. Korelacija PhysHlth i BMI

```
> # Deskriptivna statistika - numericka varijabla (BMI)
> median_bmi <- median(data$BMI)
> print(median_bmi)
[1] 29

> mean_bmi <- mean(data$BMI)
> print(mean_bmi)
[1] 29.77

> standard_deviation <- sd(data$BMI)
> print(standard_deviation)
[1] 6.806707

> variance <- var(data$BMI)
> print(variance)
[1] 46.33127

> quantiles <- quantile(data$BMI, probs = c(0.25, 0.5, 0.75))
> print(quantiles)
25% 50% 75%
 25  29  33

> correlation <- cor(data$PhysHlth, data$BMI)
> print(correlation)
[1] 0.1499764
```

summary(data) # za sve varijable

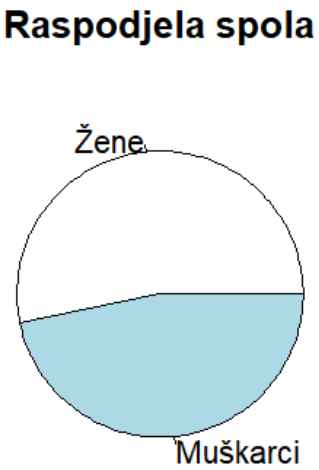
```
> summary(data)
 Diabetes_binary     HighBP          HighChol        CholCheck           BMI             Smoker          Stroke       HeartDiseaseorAttack
 1:1000          Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :14.00   Min.   :0.000   Min.   :0.000   Min.   :0.000
 2:1000          1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:25.00   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.000
                 Median :1.0000   Median :1.000   Median :1.000   Median :28.00   Median :0.000   Median :0.000   Median :0.000
                 Mean   :0.5775   Mean   :0.528   Mean   :0.972   Mean   :29.69   Mean   :0.475   Mean   :0.072   Mean   :0.156
                 3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:33.00   3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000
                 Max.   :1.0000   Max.   :1.000   Max.   :1.000   Max.   :92.00   Max.   :1.000   Max.   :1.000   Max.   :1.000
  PhysActivity        Fruits          Veggies        HvyAlcoholConsump AnyHealthcare    NoDocbcCost        GenHlth          MentHlth
 Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.000     Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   : 0.000
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:1.0000   1st Qu.:0.000     1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.: 0.000
 Median :1.000   Median :1.000   Median :1.0000   Median :0.000     Median :1.0000   Median :0.0000   Median :3.000   Median : 0.000
 Mean   :0.704   Mean   :0.605   Mean   :0.7875   Mean   :0.046     Mean   :0.9585   Mean   :0.0975   Mean   :2.824   Mean   : 3.865
 3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.000     3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.: 3.000
 Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.000     Max.   :1.0000   Max.   :1.0000   Max.   :5.000   Max.   :30.000
    PhysHlth        DiffWalk           Sex             Age           Education        Income                BMI_Group     Age_Group
 Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.000   Min.   :1.000   Min.   :1.000   Underweight   : 17   18-34 :125
 1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 7.000   1st Qu.:4.000   1st Qu.:4.000   Healthy weight:439   35-54 :280
 Median : 0.000   Median :0.0000   Median :0.0000   Median : 9.000   Median :5.000   Median :6.000   Overweight    :672   55-69 :750
 Mean   : 5.923   Mean   :0.2495   Mean   :0.4685   Mean   : 8.635   Mean   :4.932   Mean   :5.713   Obesity       :706   70-80+:701
 3rd Qu.: 5.250   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:11.000   3rd Qu.:6.000   3rd Qu.:8.000   Severe obesity:166   NA's  :144
 Max.   :30.000   Max.   :1.0000   Max.   :1.0000   Max.   :13.000   Max.   :6.000   Max.   :8.000
 Education_Group
 Min.   :1.000
 1st Qu.:2.000
 Median :3.000
 Mean   :2.634
 3rd Qu.:3.000
 Max.   :3.000
```

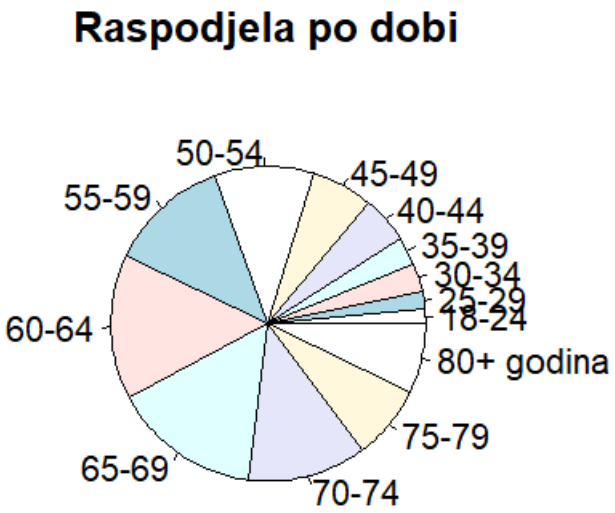
DESKRIPTIVNA STATISTIKA

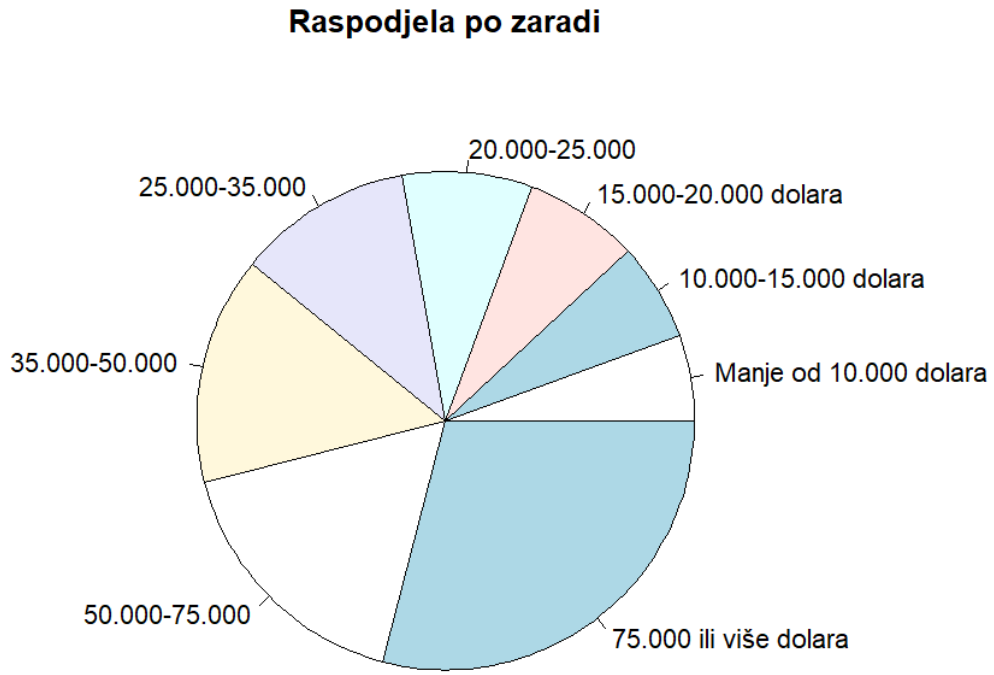


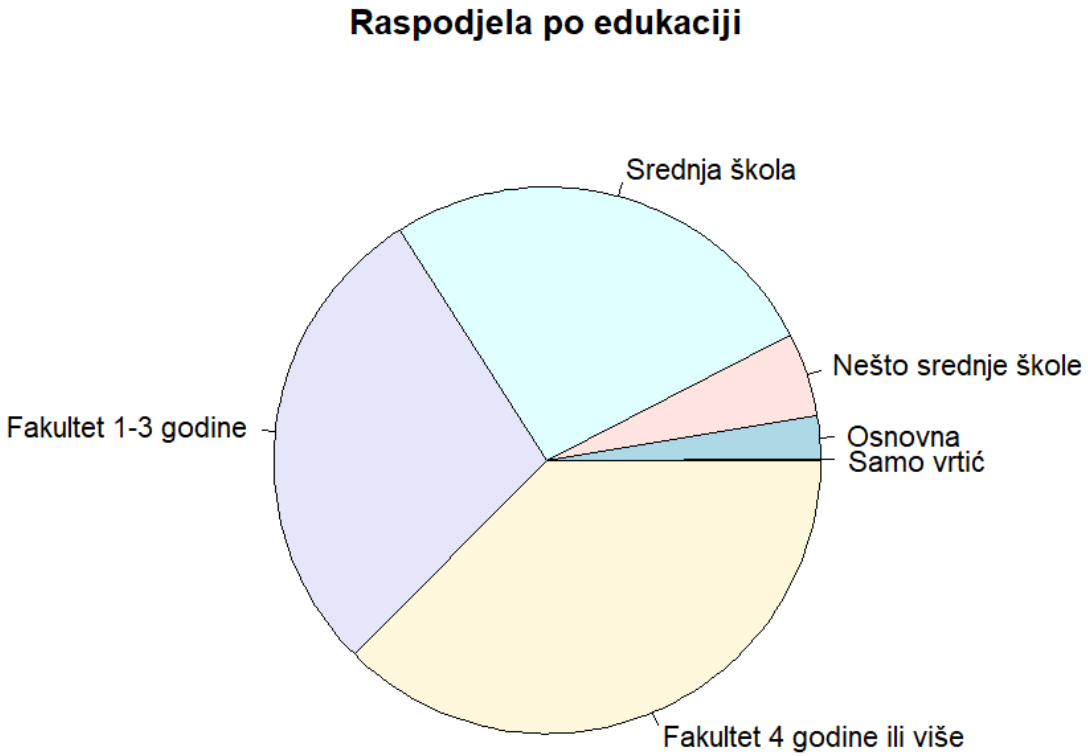
Slika 2. Raspodjela kvalitativnih varijabli

```r
# Deskriptivna statistika - kvalitativne varijable
promatranja <- table(data$Diabetes_binary)
pie(promatranja, labels = c("Nema dijabetes", "Ima dijabetes"), main =
"Raspodjela po dijabetesu")

promatranja <- table(data$Sex)
pie(promatranja, labels = c("Žene", "Muškarci"), main = "Raspodjela spola")

promatranja <- table(data$HighChol)
pie(promatranja, labels = c("Visok kolesterol", "Nizak kolesterol"),
                            main =  "Raspodjela po kolesterolu")

promatranja <- table(data$Age)
pie(promatranja, labels = c("18-24", "25-29", "30-34", "35-39", "40-44",
                            "45-49", "50-54", "55-59", "60-64", "65-69",
                            "70-74", "75-79", "80+ godina"),
    main = "Raspodjela po dobi")

promatranja <- table(data$Education)
pie(promatranja, labels = c("Samo vrtić", "Osnovna", "Nešto srednje škole",
                            "Srednja škola", "Fakultet 1-3 godine",
                            "Fakultet 4 godine ili više"),
    main = "Raspodjela po edukaciji")

promatranja <- table(data$Income)
pie(promatranja, labels = c("Manje od 10.000 dolara", "10.000-15.000 dolara",
                            "15.000-20.000 dolara", "20.000-25.000",
                            "25.000-35.000", "35.000-50.000", "50.000-75.000",
                            "75.000 ili više dolara"),
    main = "Raspodjela po zaradi")
```









Slika 2. Raspodjela kvalitativnih varijabli

PROVEDENE OBRADE

- 2 nezavisna uzorka
  - dijabetičari vs ne-dijabetičari
- Odabrane varijable (kontinuirane, kvalitativne)
  - BMI: x ε [14, 92]
  - Income: 1, 2, 3, 4, 5, 6, 7, 8
  - HighBP: 0, 1
  - GenHlth: 1, 2, 3, 4, 5

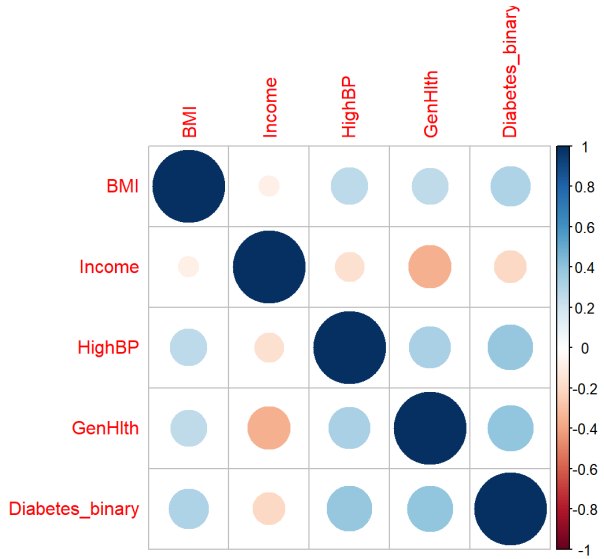
Matrica korelacije
Dijagrami rasipanja

Logistička regresija

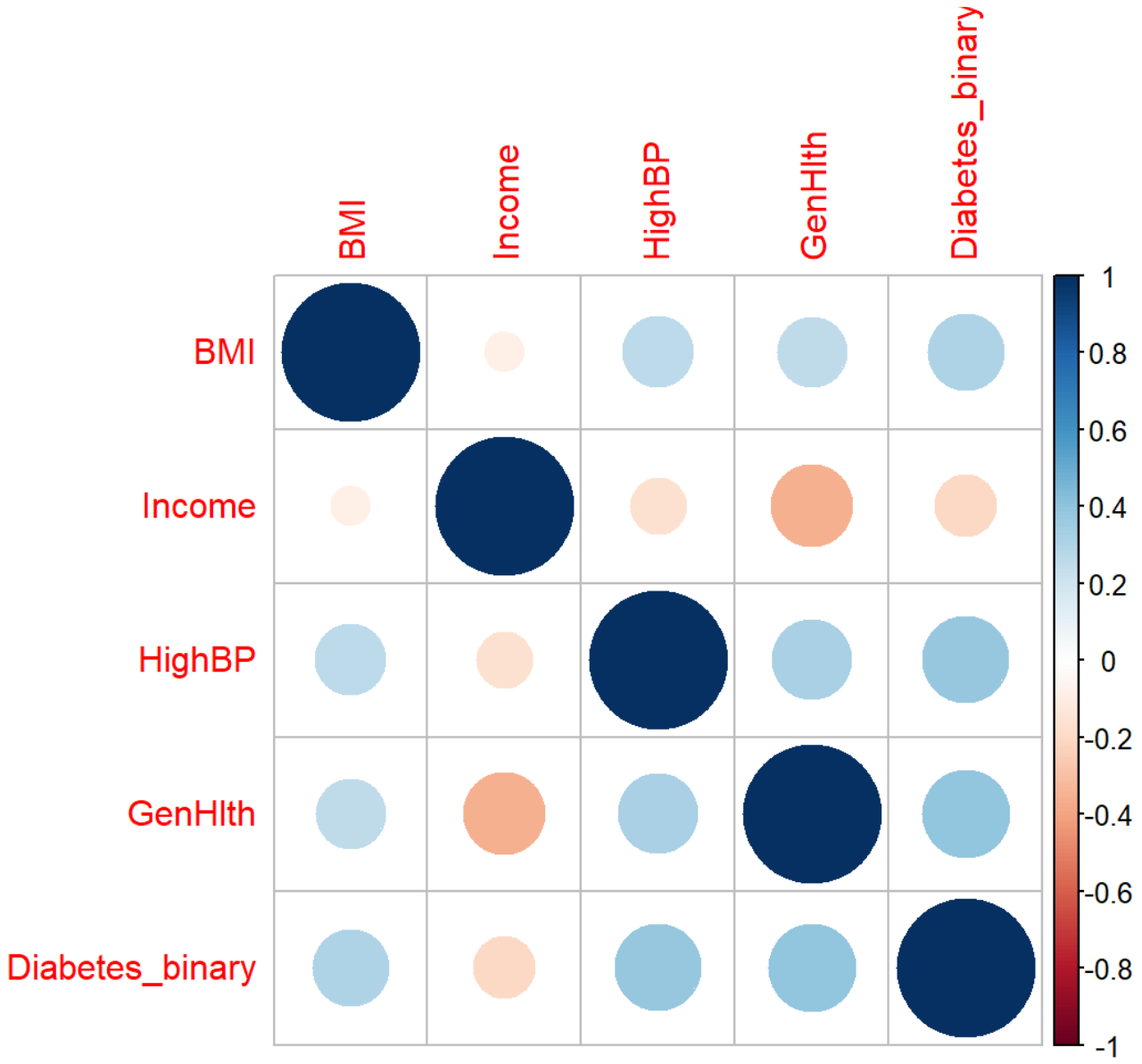
Hi-kvadrat

MATRICA KORELACIJE

```
##                      BMI      Income     HighBP    GenHlth Diabetes_binary
## BMI            1.00000000 -0.08070523  0.2653145  0.2565445       0.3015427
## Income        -0.08070523  1.00000000 -0.1666267 -0.3541305      -0.2034299
## HighBP         0.26531455 -0.16662669  1.0000000  0.3273458       0.3857488
## GenHlth        0.25654447 -0.35413051  0.3273458  1.0000000       0.3982426
## Diabetes_binary 0.30154274 -0.20342990 0.3857488  0.3982426       1.0000000
```



Slika 3. Matrica korelacije

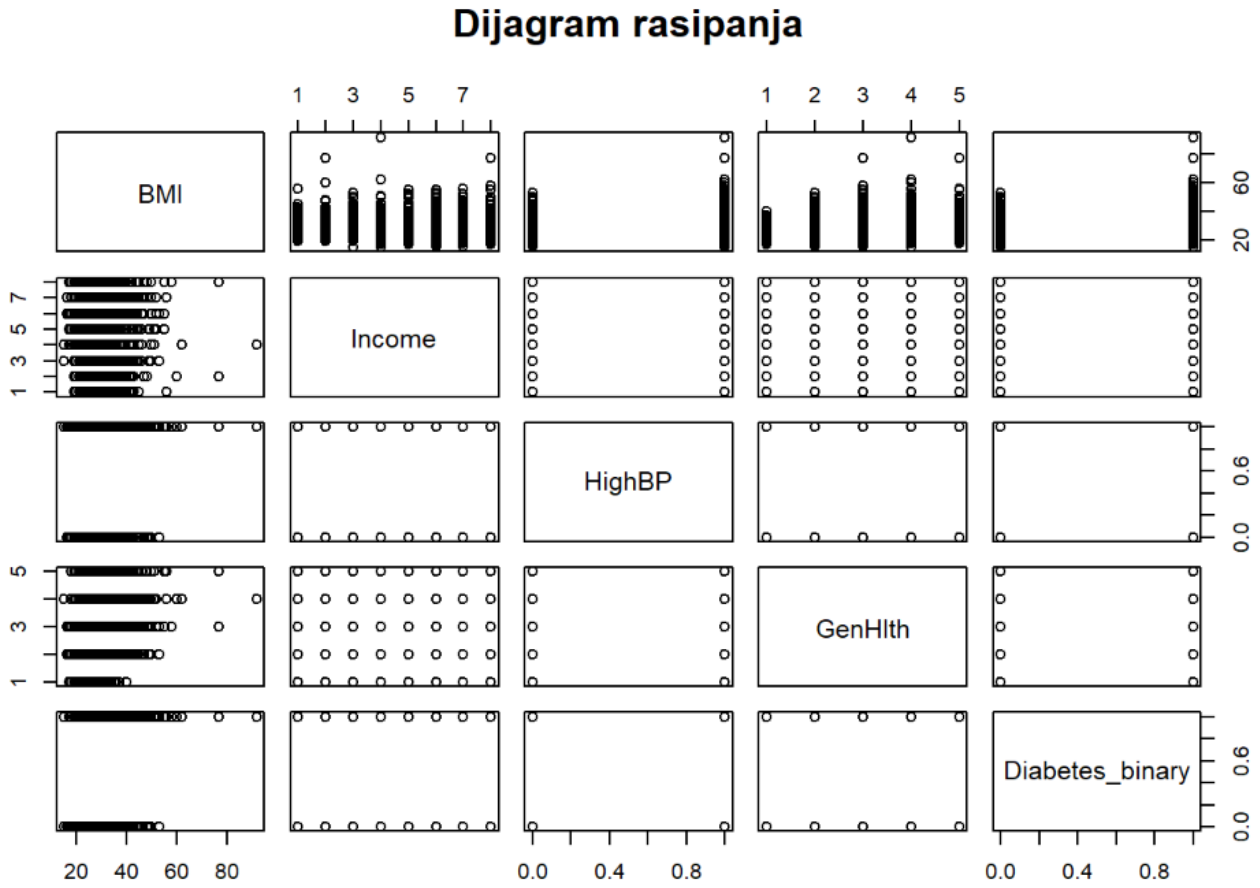


Slika 3. Matrica korelacije

```r
# matrica korelacije
correlation_matrix <-
  cor(numeric_vars, use = "complete.obs")

print(correlation_matrix)
corrplot(correlation_matrix,
         method = "circle")
```

DIJAGRAM RASIPANJA



Slika 4. Dijagram rasipanja

```
# dijagrami rasipanja
pairs(~ BMI + Income + HighBP +
  GenHlth + Diabetes_binary,
    main = "Dijagram rasipanja",
    data = data)
```

LOGISTIČKA REGRESIJA FORMULA

$$logit(DiabetesBinary) = \beta_0 + \beta_1 BMI + \beta_2 HighBP + \beta_3 GenHlth + \beta_4 Income + \epsilon$$

$$ln(P(imati\_dijabetes)/P(nemati\_dijabetes))$$

PROVEDBA

```
data$Diabetes_binary <- as.factor(data$Diabetes_binary)
data$Income <- as.factor(data$Income)
data$HighBP <- as.factor(data$HighBP)
data$GenHlth <- as.factor(data$GenHlth)

skup <- data %>%
  select(Diabetes_binary, HighBP, GenHlth, Income, BMI)

set.seed(1)
split = initial_split(skup, prop = 0.7, strata = Diabetes_binary)

train = split %>%
  training()
test = split %>%
  testing()
```

OBRADA LOGISTIČKA REGRESIJA

```
attach(train)
summary(train)
```

```
##  Diabetes_binary HighBP  GenHlth      Income          BMI
##  0:700           0:598   1:142   8       :385   Min.   :15.00
##  1:700           1:802   2:432   7       :224   1st Qu.:25.00
##                          3:449   6       :216   Median :29.00
##                          4:263   5       :169   Mean   :29.65
##                          5:114   4       :157   3rd Qu.:33.00
##                                  3       :114   Max.   :92.00
##                                  (Other):135
```

OBRADA LOGISTIČKA REGRESIJA

```
describe(train)
```

```
##                    vars    n  mean   sd median trimmed  mad min max range  skew kurtosis   se
## Diabetes_binary*      1 1400  1.50 0.50    1.5    1.50 0.74   1   2     1  0.00    -2.00 0.01
## HighBP*               2 1400  1.57 0.49    2.0    1.59 0.00   1   2     1 -0.29    -1.91 0.01
## GenHlth*              3 1400  2.84 1.10    3.0    2.82 1.48   1   5     4  0.23    -0.63 0.03
## Income*               4 1400  5.69 2.10    6.0    5.91 2.97   1   8     7 -0.60    -0.71 0.06
## BMI                   5 1400 29.65 6.72   29.0   29.04 5.93  15  92    77  1.62     7.76 0.18
```

REZULTATI LOGISTIČKA REGRESIJA

```
model <- glm(Diabetes_binary ~ BMI + Income + HighBP + GenHlth,
  data = train, family = binomial)
summary(model)
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.36003    0.54125  -9.903  < 2e-16 ***
## BMI          0.06975    0.01115   6.254 3.99e-10 ***
## Income2      0.78698    0.40396   1.948 0.051398 .
## Income3      1.02183    0.37235   2.744 0.006065 **
## Income4      0.76518    0.34767   2.201 0.027743 *
## Income5      1.18561    0.35285   3.360 0.000779 ***
## Income6      0.61090    0.33848   1.805 0.071098 .
## Income7      0.70241    0.33737   2.082 0.037341 *
## Income8      0.38522    0.32775   1.175 0.239850
## HighBP1      1.20292    0.13132   9.160  < 2e-16 ***
## GenHlth2     1.34604    0.33071   4.070 4.70e-05 ***
## GenHlth3     2.37584    0.32999   7.200 6.03e-13 ***
## GenHlth4     2.51745    0.34872   7.219 5.23e-13 ***
## GenHlth5     2.80037    0.39705   7.053 1.75e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1940.8  on 1399  degrees of freedom
## Residual deviance: 1488.9  on 1386  degrees of freedom
## AIC: 1516.9
##
## Number of Fisher Scoring iterations: 5
```

REZULTATI LOGISTIČKA REGRESIJA

$$\begin{aligned} logit(DiabetesBinary) = &-5.36003 + 0.06975 * BMI + 1.20292 * HighBP1 + 1.34604GenHlth2 \\ &+ 2.37584GenHlth3 + 2.51745GenHlth4 + 2.80037GenHlth5 + 0.78698Income2 + 1.02183Income3 \\ &+ 0.76518Income4 + 1.18561Income5 + 0.61090Income6 + 0.70241Income7 + 0.38522Income8 \end{aligned}$$

KOEFICIJENTI LOGISTIČKE REGRESIJE & OMJERA IZGLEDA

```
exp(model$coefficients)
```

```
##  (Intercept)          BMI      Income2      Income3      Income4      Income5      Income6      Income7      Income8
##  0.004700749  1.072236345  2.196753842  2.778278951  2.149387700  3.272666611  1.842095207  2.018606202  1.469943960
##      HighBP1     GenHlth2     GenHlth3     GenHlth4     GenHlth5
##  3.329819837  3.842188376 10.760003655 12.396899569 16.450794905
```

KOEFICIJENTI LOGISTIČKE REGRESIJE & OMJERA IZGLEDA

```
##                    2.5 %      97.5 %
## (Intercept) 0.001581672  0.01325646
## BMI         1.049385430  1.09629860
## Income2     0.999029915  4.88480843
## Income3     1.341095643  5.79091611
## Income4     1.086509315  4.25935115
## Income5     1.640188370  6.56144053
## Income6     0.947701662  3.58412113
## Income7     1.040968740  3.91996879
## Income8     0.772324534  2.80130492
## HighBP1     2.577079346  4.31311992
## GenHlth2    2.081990516  7.69080889
## GenHlth3    5.842622516 21.51772115
## GenHlth4    6.465416517 25.60844677
## GenHlth5    7.781523726 37.16838503
```

```
exp(confint(model))
```

SIGNIFIKANTNOST MODELA

- logistički regresijski model je statistički značajan na razini od 5%
- model je bolji od nultog modela

```
# testna statistika
ModelChi = model$null.deviance
  - model$deviance
ModelChi
```

```
[1] 451.9521
```

```
# broj stupnjeva slobode
ChiDf = model$df.null - model$df.residual
ChiDf
```

```
[1] 13
```

```
# p-vrijednost
ChisqProb = 1 - pchisq(ModelChi, ChiDf)
ChisqProb
```

```
[1] 0
```

ZAKLJUČAK

- Varijable `BMI + Income + HighBP + GenHlth` u logističkoj regresiji modeliraju statistički značajnu korelaciju

regresija.Rmd

GitHub poveznica na ovu datoteku u trenutku izrade dokumentacije.

```
---
title: "Regresija"
output: html_document
---

-----------------------------------------------------------------------

# 2. faza projekta - Logistička regresija

```{r}
if (!require("dplyr")) install.packages("dplyr")

library(ggplot2)
library(corrplot)

data <- read.csv(file.choose())

# matrica korelacije
correlation_matrix <- cor(numeric_vars, use = "complete.obs")
print(correlation_matrix)
corrplot(correlation_matrix, method = "circle")

# dijagrami rasipanja
pairs(~ BMI + Income + HighBP + GenHlth + Diabetes_binary,
  main = "Dijagram rasipanja", data = data)
```

Jednadžba:

$$logit(DiabetesBinary) = \beta_{0} + \beta_1BMI + \beta_2HighBP + \beta_3GenHlth +
\beta_4Income + \epsilon$$

$logit(admit)$ je $ln(P(imati\_dijabetes)/P(nemati\_dijabetes))$

```{r}
data$Diabetes_binary <- as.factor(data$Diabetes_binary)
data$Income <- as.factor(data$Income)
data$HighBP <- as.factor(data$HighBP)
data$GenHlth <- as.factor(data$GenHlth)


if (!require("rsample")) install.packages("rsample")
library(rsample)

skup <- data %>%
  select(Diabetes_binary, HighBP, GenHlth, Income, BMI)

set.seed(1)
split = initial_split(skup, prop = 0.7, strata = Diabetes_binary)
train = split %>%
  training()
test = split %>%
  testing()

attach(train)
summary(train)

library(psych)
describe(train)

model <- glm(Diabetes_binary ~ BMI + Income + HighBP + GenHlth, data = train, family = binomial)
summary(model)
```

$$logit(DiabetesBinary) = -5.36003 + 0.06975*BMI + 1.20292*HighBP1 + 1.34604GenHlth2 + 2.37584GenHlth3 + 2.51745GenHlth4 +
2.80037GenHlth5 + 0.78698Income2 + 1.02183Income3 + 0.76518Income4 + 1.18561Income5 + 0.61090Income6 + 0.70241Income7 +
0.38522Income8$$

Koeficijenti logističke regresije i omjera izgleda

```{r}
exp(model$coefficients)
exp(confint(model))
```

Signifikantnost modela

```{r}
# testna statistika
ModelChi = model$null.deviance - model$deviance
ModelChi

# broj stupnjeva slobode
ChiDf = model$df.null - model$df.residual
ChiDf

# p-vrijednost
ChisqProb = 1 - pchisq(ModelChi, ChiDf)
ChisqProb
```
```

LITERATURA

Sav programski kod se može pronaći na javnom GitHub repozitoriju ovog rada: https://github.com/jfletcher20/diabetes-health-indicators-analysis

- *AGE5GYR - Variable Home Page* (bez dat.). Preuzeto na: https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/AGEG5YR?archive=NAHDAP (pristupano: 5.5.2024.).
- *Body mass index (BMI) | NHS inform* (bez dat.). Preuzeto na: https://www.nhsinform.scot/healthy-living/food-and-nutrition/healthy-eating-and-weight-loss/body-mass-index-bmi/ (pristupano: 5.5.2024.).
- *Diabetes Health Indicators Dataset* (bez dat.). Preuzeto na: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv (pristupano: 5.5.2024.).
- *EDUCA - Variable Home Page* (bez dat.). Preuzeto na: https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/EDUCA?archive=nahdap (pristupano: 5.6.2024.).
- *INCOME2 - Variable Home Page* (bez dat.). Preuzeto na: https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/INCOME2?archive=NAHDAP (pristupano: 5.6.2024.).

HVALA VAM NA PAŽNJI

JOSHUA LEE FLETCHER, NOA MIDŽIĆ