PROJEKT GRUPE KOMETA

Josip Torić, Matej Pipalović, Patrik Marić

May 24, 2018

Contents

1 UVOD **2**

2 UČITAVANJE PODATAKA **3**

3 DESKRIPTIVNA STATISTIKA **4**

4 INFERENCIJALNA **9**
4.1 t-test 9
4.2 ANOVA 13

5 STROJNO UČENJE **14**
5.1 Linearna regresija 14
5.2 Logistička regresija 19
5.3 Stroj potpornih vektora 22

6 ZAKLJUČAK **24**

LITERATURA **25**

1 UVOD

Kažu da su ljudi napredna bića. Kažu da ljudi su različiti od životinja. Kažu da se ljudi vode razumom, a ne nagonima. No, neki vrlo jednostavni eksperimenti kao što je ga je na Stanfordu providio Phillip George Zimbardo govore suprotno. Zimbardo u svojem pokusu opisanom u (Zimbardo 1971) je studente podijelio u svije skupine, zatvorenike i čuvare te ih smjestio u zatvor. Nakon nekog vremena čuvari su postali bahati i okrutni prema zatvorenicima, a zatvorenici su nakon nekog vremena digli bunu i ozbiljno se potukli s čuvarima. Dakle, ljudi su se ponašali kao da se nikad nisu poznavali i bili okrutni jedni prema drugima samo zato što su se našli u takvim pozicijama.

U ovom radu nećemo govoriti o zatvorenicima, već o studentima FER-a, a umjesto Zimbarda u glavnoj ulozi je prof.dr.sc. Mile Šikić. Profesor Šikić nije podijelio studente u zatvorenike i čuvare, već je pred studente postavio jednu knjigu te ih pitao da estimiraju koliko knjiga ima stranica. Također, pitao je učenike da estimiraju koliko će njihove kolege reći koliko knjiga ima stranica, te koliko će njihovi prijatelji reći. Zatim je ponovio eksperiment te rekao da trojica koji budu najbliži će dobiti 3 boda iz bodova s predavanja. Iz podataka koje je prof. Šikić prikupio pokušat ćemo vidjeti da li studenti FER-a bolje razmišljanju kada je nagrada u pitanju. Stvaran broj stranica knjiga je 1171.

Prvo ćemo učitati dataset, zatim ćemo prvo analizirati uz pomoć deskriptivne statistike, izvući ćemo sumarizacije podataka i iscrtati najvažnije grafove. Zatim ćemo uz pomoć inferencijalne analize pokušati izvući zaključke o podacima te na kraju pokušati primijeniti strojno učenje na podatke. Naposljetku donesen je zaključak o svemu što smo napravili.

2 UČITAVANJE PODATAKA

Učitat ćemo dani dataset iz .csv filea.

```
podaci <- tbl_df(fread("sap.csv",header=T))
```

Iz podataka nam nisu bitni podaci o grupi i rednom broju studenta jer iz njih ne možemo ništa zaključiti. Grupe su neravnomjerno raspoređene te je 80 % studenata iz grupe P01, a ostale grupe su zastupljene od 3-5 %. Redni broj studenta je prilično jasan zašto nije bitan za daljnju analizu. Također, pretvorili smo spol u faktorsku varijablu da bi je mogli lakše kasnije analizirati.

```
podaci %>% dplyr::select(-GRUPA,-Br.stud) -> podaci

podaci$SPOL <- as.factor(podaci$SPOL)
```

Stvorit ćemo i nekoliko pomoćnih datasetova. Jedan od njih koji će služiti da vidimo kako naga+rada utječe na procjenu je kada predikcije bez obzira na nagradu stavimo u jednu tablicu te dodamo još jedan stupac koji će nam govoriti je li bila nagrada ili nije. Također ćemo razdvojiti dataset po spolu i pripremiti podatke za linearnu regresiju.

```
podaci %>% dplyr::select(student,cijela_grupa,samo_prijatelji,MI,SPOL) -> temp1
temp1$nagrada <- FALSE
podaci %>% dplyr::select(student=nagrada_student,
                         cijela_grupa=nagrada_cijela_grupa,
                         samo_prijatelji=nagrada_samo_prijatelji,MI,SPOL) -> temp2
temp2$nagrada <- TRUE

podaciNagrada <- rbind(temp1,temp2)


podaci %>% dplyr::select(-SPOL) -> podaciLinear


podaci %>% filter(podaci$SPOL == "M") -> muski
podaci %>% filter(podaci$SPOL == "F") -> zenski
```

3 DESKRIPTIVNA STATISTIKA

Deskriptivna statistička analiza bavi se mjerama centralne tendencije i mjerama rasipanja. Za razliku od inferencijalne statističke analize, ona ne počiva na teoriji vjerojatnosti. Deskriptivna statistička analiza nam daje jednostavne zaključke o uzorku. Pregled deskriptive analize napravljen je prema (Diez, Barr, and Cetinkaya-Rundel 2015).

Najvažnije mjere centralne tendencije su aritmetička sredina, medijan i mod.

U simetričkim distribucijama, mod, medijan i srednja vrijednost su jednaki, dok u zakrivljenim distribucijama srednja vrijednost teži u smjeru zakrivljenosti distribucije. U slučaju zakrivljenih distribucija, medijan je najbolji pokazatelj sredine.

Najvažnije mjere rasipanja su rang, interkvartilni rang, varijanca, standardna devijacija, koeficijent varijacije.

```
summary(podaci)
```

```
##     student        cijela_grupa  samo_prijatelji  nagrada_student
##  Min.   : 350.0   Min.   : 200   Min.   : 234.0   Min.   : 343.0
##  1st Qu.: 485.2   1st Qu.: 500   1st Qu.: 470.0   1st Qu.: 493.8
##  Median : 600.0   Median : 600   Median : 600.0   Median : 590.0
##  Mean   : 652.4   Mean   : 601   Mean   : 625.5   Mean   : 651.5
##  3rd Qu.: 750.0   3rd Qu.: 700   3rd Qu.: 750.0   3rd Qu.: 750.0
##  Max.   :1600.0   Max.   :1139   Max.   :1200.0   Max.   :1300.0
##  nagrada_cijela_grupa nagrada_samo_prijatelji       MI        SPOL
##  Min.   : 300.0       Min.   : 250.0          Min.   : 6.50   F:22
##  1st Qu.: 500.0       1st Qu.: 500.0          1st Qu.:15.00   M:78
##  Median : 600.0       Median : 585.0          Median :18.75
##  Mean   : 639.4       Mean   : 625.4          Mean   :18.38
##  3rd Qu.: 713.2       3rd Qu.: 750.0          3rd Qu.:22.12
##  Max.   :1200.0       Max.   :1020.0          Max.   :28.00
```

Iz osnovnih mjera o podatcima vidimo da predikcije studenata nisu ni blizu stvarnoj brojci stranica knjige koji iznosi 1171. Također, možemo vidjeti da studenti bolje procjenjuju kako će grupa procjenjivati kada je u pitanju nagrada. Isto tako vidimo da na GER-u ima dosta više muškaraca nego žena.

```
muski %>% summary
```

```
##     student        cijela_grupa    samo_prijatelji  nagrada_student
##  Min.   : 383.0   Min.   : 200.0   Min.   : 234.0   Min.   : 378.0
##  1st Qu.: 500.0   1st Qu.: 500.0   1st Qu.: 470.0   1st Qu.: 522.8
##  Median : 616.0   Median : 600.0   Median : 600.0   Median : 624.0
##  Mean   : 668.5   Mean   : 606.6   Mean   : 634.8   Mean   : 665.4
##  3rd Qu.: 792.0   3rd Qu.: 700.0   3rd Qu.: 750.0   3rd Qu.: 763.5
##  Max.   :1600.0   Max.   :1139.0   Max.   :1200.0   Max.   :1250.0
##  nagrada_cijela_grupa nagrada_samo_prijatelji       MI        SPOL
##  Min.   : 300.0       Min.   : 250.0          Min.   : 6.50   F: 0
##  1st Qu.: 500.0       1st Qu.: 500.0          1st Qu.:16.12   M:78
##  Median : 600.0       Median : 600.0          Median :19.75
##  Mean   : 642.1       Mean   : 639.6          Mean   :18.94
##  3rd Qu.: 724.0       3rd Qu.: 757.8          3rd Qu.:22.88
##  Max.   :1200.0       Max.   :1020.0          Max.   :28.00
```

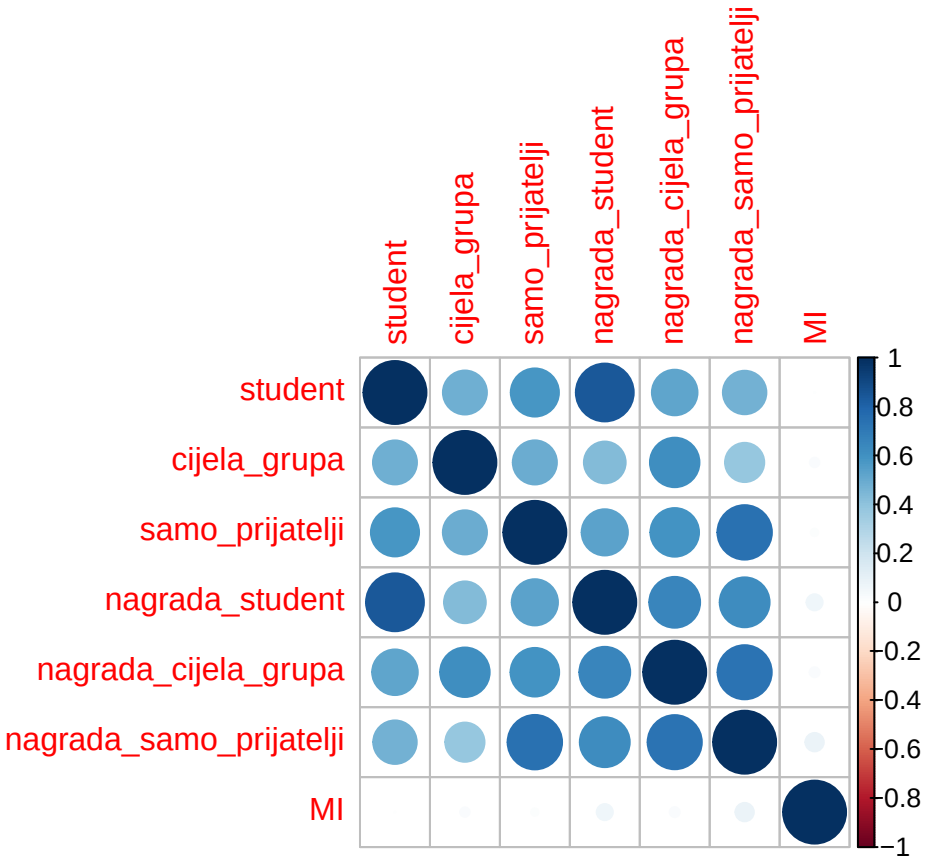
```
zenski %>% summary
```

```
##     student        cijela_grupa    samo_prijatelji  nagrada_student
##  Min.   : 350.0   Min.   : 300.0   Min.   : 350.0   Min.   : 343.0
##  1st Qu.: 467.2   1st Qu.: 500.0   1st Qu.: 500.0   1st Qu.: 459.5
```

```
##  Median : 550.0   Median : 600.0   Median : 540.0   Median : 529.5
##  Mean   : 595.5   Mean   : 581.0   Mean   : 592.7   Mean   : 602.1
##  3rd Qu.: 622.5   3rd Qu.: 682.5   3rd Qu.: 700.0   3rd Qu.: 664.2
##  Max.   :1200.0   Max.   :1000.0   Max.   :1005.0   Max.   :1300.0
##  nagrada_cijela_grupa nagrada_samo_prijatelji       MI        SPOL
##  Min.   : 312.0       Min.   : 300.0          Min.   :10.00   F:22
##  1st Qu.: 515.5       1st Qu.: 500.0          1st Qu.:13.50   M: 0
##  Median : 600.0       Median : 522.0          Median :16.00
##  Mean   : 630.0       Mean   : 574.9          Mean   :16.39
##  3rd Qu.: 700.0       3rd Qu.: 615.0          3rd Qu.:19.00
##  Max.   :1110.0       Max.   :1000.0          Max.   :23.00
```

Iz razlike između sumarizacija muških i ženskih predikcija, vidimo da žene u načelu kažu da je broj stranica knjige manji od muškaraca te da također postižu lošije rezultate na međuispitu.

```
podaciLinear %>% cor %>% corrplot
```

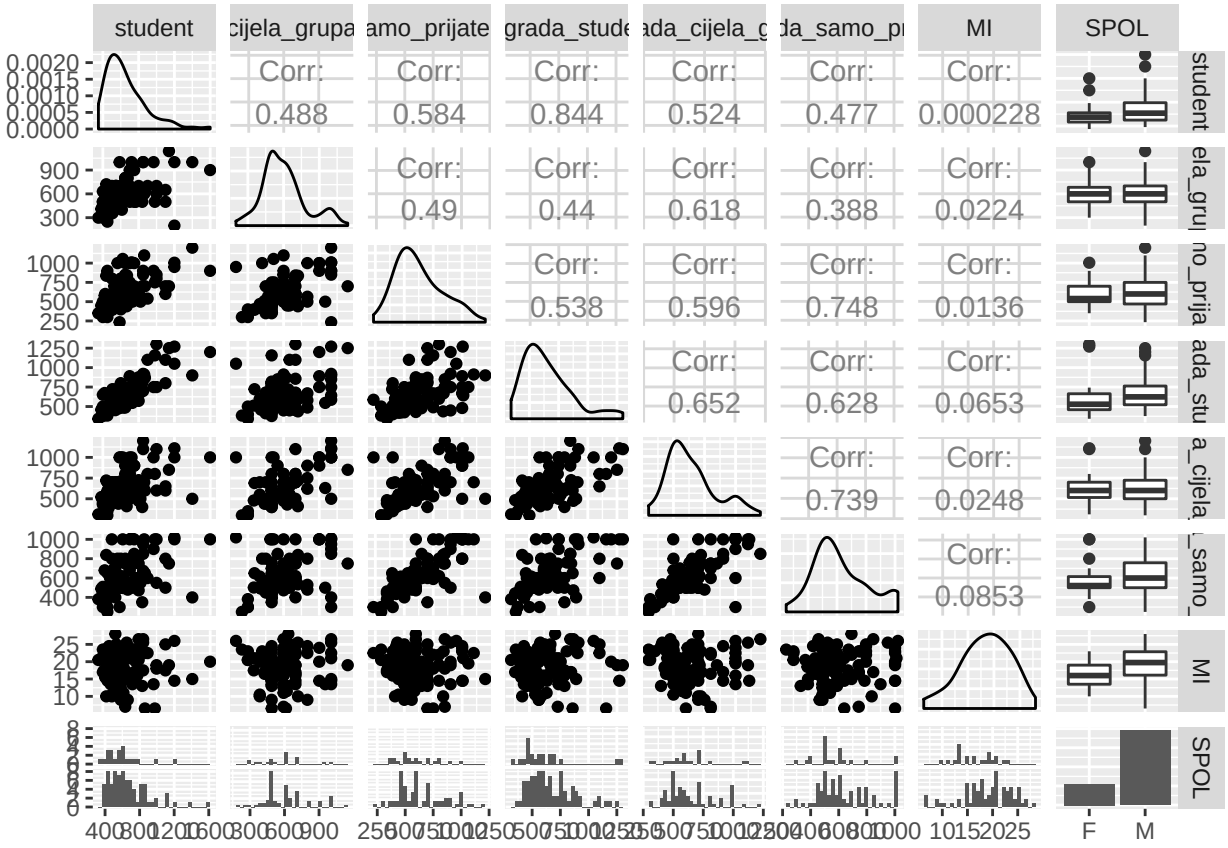


Iz matrica korelacija vidimo da rezultati međuispita ne ovise gotovo o ničemu, dok ostale varijable imaju veliku međusobnu korelaciju.

```
ggpairs(podaci)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

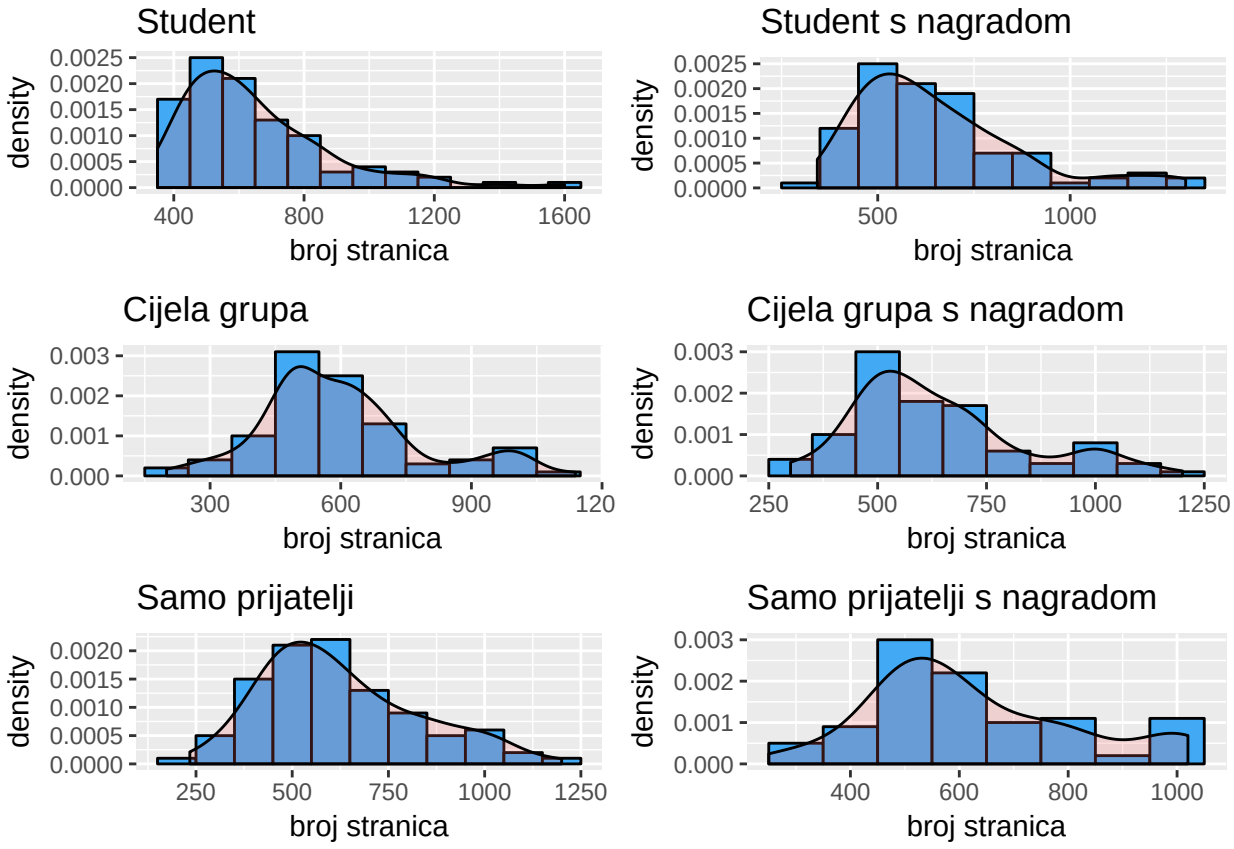
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Iz ovih skupa grafova vidimo da podaci ne podilaze savršeno normalnoj razdiobi. Iz boxplotova po spolu vidimo da samo rezultati međuispita ovise donekle o spolu.

```
my_histo_with_density <- function(data, parameter, hist_title, bindwidth, label_x) {
  graph <- data %>% ggplot(aes(parameter))+
    geom_histogram(aes(y=..density..), col="black",
                   fill="#42aaf4", binwidth = bindwidth)+
    labs(title=hist_title, x=label_x)+
    geom_density(alpha=.2, fill="#FF6666")

  return(graph)
}

bindwidth <- 100
label_x <- "broj stranica"

podaci %>% filter(!is.na(podaci$student)) -> stud_pod1
sh <- my_histo_with_density(stud_pod1, stud_pod1$student,
                            "Student", bindwidth, label_x)

podaci %>% filter(!is.na(podaci$cijela_grupa)) -> stud_pod2
cgh <- my_histo_with_density(stud_pod2, stud_pod2$cijela_grupa,
                             "Cijela grupa", bindwidth, label_x)

podaci %>% filter(!is.na(podaci$samo_prijatelji)) -> stud_pod3
sph <- my_histo_with_density(stud_pod3, stud_pod3$samo_prijatelji,
```

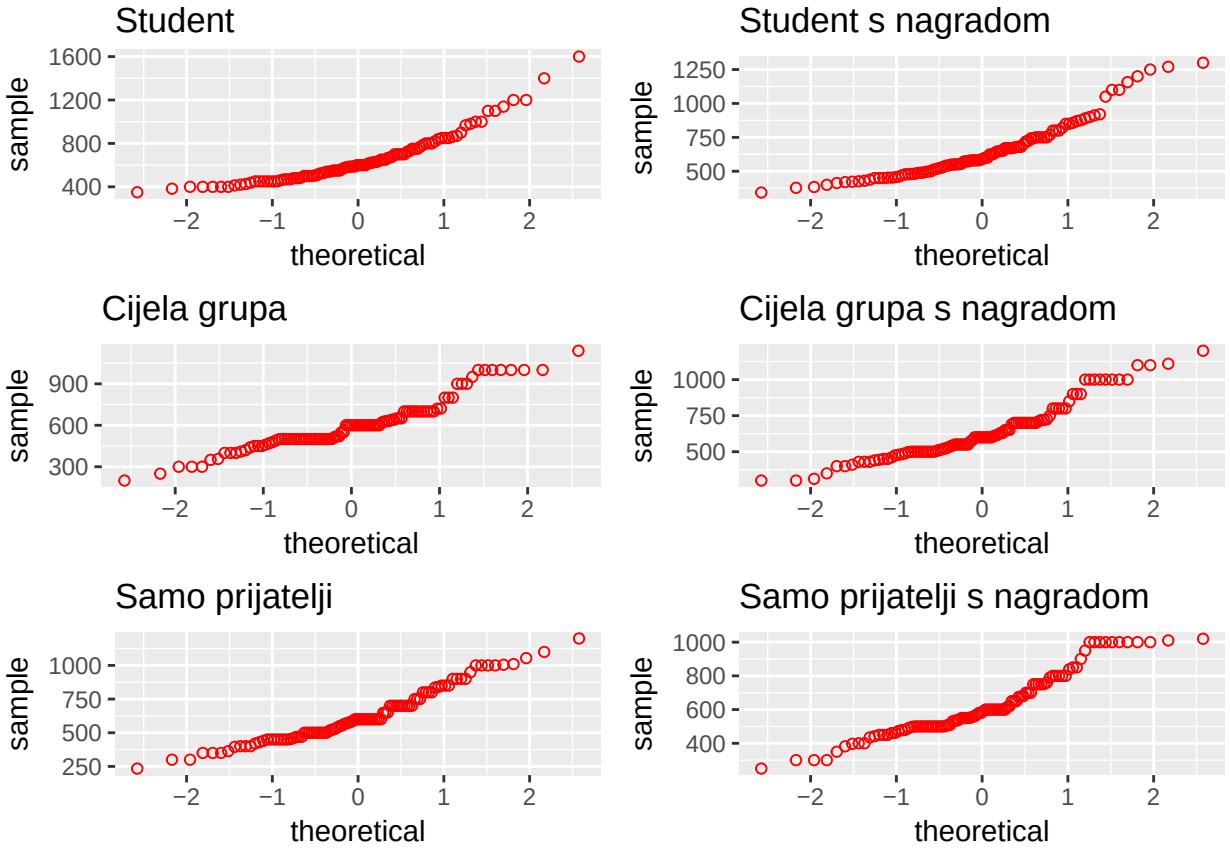
```
                              "Samo prijatelji", bindwidth, label_x)

podaci %>% filter(!is.na(podaci$nagrada_student)) -> stud_pod4
nsh <- my_histo_with_density(stud_pod4, stud_pod4$nagrada_student,
                              "Student s nagradom", bindwidth, label_x)

podaci %>% filter(!is.na(podaci$nagrada_cijela_grupa)) -> stud_pod5
ncgh <- my_histo_with_density(stud_pod5, stud_pod5$nagrada_cijela_grupa,
                              "Cijela grupa s nagradom", bindwidth, label_x)

podaci %>% filter(!is.na(podaci$nagrada_samo_prijatelji)) -> stud_pod6
nsph <- my_histo_with_density(stud_pod6, stud_pod6$nagrada_samo_prijatelji,
                              "Samo prijatelji s nagradom", bindwidth, label_x)

grid.arrange(sh, nsh, cgh, ncgh, sph, nsph, ncol=2, nrow = 3)
```



Iz ovih histograma vidimo ljepši prikaz već prikaz već nacrtanih histograma koje smo nacrtali uz pomoć funkcije ggpairs.

```
my_qq_plot <- function(data, parameter, title) {
  data %>% ggplot(aes(sample=parameter))+
    stat_qq(col="red", shape=1)+
    labs(title=title)
}

student_qq_plot <- my_qq_plot(stud_pod1,
                              stud_pod1$student, "Student")
```

```
cijela_grupa_qq_plot <- my_qq_plot(stud_pod2,                                    stud_pod2$cijela_grupa

samo_prijatelji_qq_plot <- my_qq_plot(stud_pod3,                                      stud_pod3$samo_pri

student_qq_plot_nagrada <- my_qq_plot(stud_pod4,                                      stud_pod4$nagrada_

cijela_grupa_qq_plot_nagrada <- my_qq_plot(stud_pod5,                                           stud_poc

samo_prijatelji_qq_plot_nagrad <- my_qq_plot(stud_pod6,                                            stuc

grid.arrange(student_qq_plot, student_qq_plot_nagrada,
             cijela_grupa_qq_plot, cijela_grupa_qq_plot_nagrada,
             samo_prijatelji_qq_plot, samo_prijatelji_qq_plot_nagrad, ncol=2, nrow=3)
```



Iz Q-Q plota možemo najbolje vidjeti da podaci ne podilaze normalnoj razdiobi.

4 INFERENCIJALNA

Inferencijalna analiza odnosi se na provjeravanje postavljenih hipoteza uz pomoć statističkih testova. Pregled inferencijalne statističke analize napravljen je prema (Diez, Barr, and Cetinkaya-Rundel 2015).

Inferencijalna statistička analiza se provodi u nekoliko koraka.

Prvi korak je postaviti hipoteze. Kod inferencijalne statističke analize imamo dvije hipoteze. Prva hipoteza, H_0 predstavlja nul hipotezu. Nul hipoteza je hipoteza da ne postoji nikakva značajna razlika između populacija. Druga hipoteza, H_1 predstavlja alternativnu hipotezu. Alternativna hipoteza predstavlja alternativu nultoj hipotezi.

Drugi korak je odabrati testnu statistiku. Testna statistika je statistika na temelju čijih se vrijednosti donosi odluka o odbacivanju ili ne odbacivanju zadane osnove statističke hipoteze u korist njezine alternative. Postoji više vrsta testnih statistika, ovisno što želimo testirati u varijablama.

Treći korak je odabrati nivo značajnosti α ispod koje ćemo odbaciti nul hipotezu. Najčešće korištene vrijednosti za nivo značajnosti su 1% i 5%.

Četvrti korak je izračunati vrijednost statistike i usporediti ga s α. Ako je vrijednost statistike manja of α, zaključujemo da odbacujemo nul hipotezu u korist alternativne hipoteze.

Inferencijalna statistička analiza se smije provoditi samo za distribucije koje odgovaraju normalnoj distribuciji. Ovo se na prvi pogled čini kao prilično limitirajući faktor, no prema centralnog graničnom teoremu, sve distribucije teže u normalnu distribuciju. Centralni granični teorem kaže ako je $\bar{X}$ aritmetička srednja vrijednost slučajnog uzorka veličine n uzetog iz populacije s s očekivanjem μ i varijancom σ^2 onda normirana suma teži po distribuciju u normalnu distribuciju kada n teži u beskonačnost.

Iako naši podaci ne podilaze normalnoj razdiobi, zbog CGT-a možemo raditi testiranja.

Za naše podatke prvo ćemo napraviti t-test, a kasnije ANOVA test.

4.1 t-test

```
t.test(podaci$student, mu = 1171, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  podaci$student
## t = -22.683, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 1171
## 95 percent confidence interval:
##  607.078 697.802
## sample estimates:
## mean of x
##    652.44
```

Iz ovog testa vidimo da mudrost mase koja će točno predvidjeti broj stranica nema veze s vezom u ovom našem slučaju.

```
t.test(podaci$student, podaci$cijela_grupa, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  podaci$student and podaci$cijela_grupa
## t = 1.7544, df = 189.22, p-value = 0.08097
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -6.397132 109.297132
## sample estimates:
## mean of x mean of y
##    652.44    600.99
```

```
t.test(podaci$nagrada_student, podaci$nagrada_cijela_grupa, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  podaci$nagrada_student and podaci$nagrada_cijela_grupa
## t = 0.41847, df = 197.2, p-value = 0.6761
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -44.73586  68.83586
## sample estimates:
## mean of x mean of y
##    651.47    639.42
```

```
mean_stud <- mean(podaci$student)

razlike_stud_cijela_grupa <- podaci$cijela_grupa - mean_stud
t.test(razlike_stud_cijela_grupa, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  razlike_stud_cijela_grupa
## t = -2.8012, df = 99, p-value = 0.006124
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  -99.689223  -3.210777
## sample estimates:
## mean of x
##    -51.45
```

```
mean_stud_nagrada <- mean(podaci$nagrada_student)

razlike_stud_cijela_grupa_nagrad <- podaci$nagrada_cijela_grupa - mean_stud_nagrada
t.test(razlike_stud_cijela_grupa_nagrad, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  razlike_stud_cijela_grupa_nagrad
## t = -0.61165, df = 99, p-value = 0.5422
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  -63.79238  39.69238
## sample estimates:
## mean of x
##    -12.05
```

Iz ovih t-testova možemo naslutiti da ako je u pitanju nagrada, studenti će bolje procjenjivati koliko će cijela grupa predvidjeti.

```r
t.test(podaci$student, podaci$nagrada_student, paired = TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  podaci$student and podaci$nagrada_student
## t = 0.078428, df = 99, p-value = 0.9376
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.57089  25.51089
## sample estimates:
## mean of the differences
##                    0.97
```

Iz ovog testa vidimo da nagrada nema utjecaj na procjenu samog studenta.

```r
t.test(muski$student, zenski$student, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  muski$student and zenski$student
## t = 1.4681, df = 39.548, p-value = 0.15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.5532 173.6698
## sample estimates:
## mean of x mean of y
##  668.5128  595.4545
```

```r
t.test(muski$cijela_grupa, zenski$cijela_grupa, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  muski$cijela_grupa and zenski$cijela_grupa
## t = 0.59217, df = 35.164, p-value = 0.5575
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -62.21637 113.47278
## sample estimates:
## mean of x mean of y
##  606.6282  581.0000
```

```r
t.test(muski$samo_prijatelji, zenski$samo_prijatelji, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  muski$samo_prijatelji and zenski$samo_prijatelji
## t = 0.9436, df = 38.876, p-value = 0.3512
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -48.13971 132.31454
## sample estimates:
## mean of x mean of y
```

```
##  634.7692  592.6818
```

```
t.test(muski$student, muski$nagrada_student, paired = TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  muski$student and muski$nagrada_student
## t = 0.21022, df = 77, p-value = 0.8341
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -26.50302  32.75943
## sample estimates:
## mean of the differences
##                3.128205
```

```
t.test(zenski$student, zenski$nagrada_student, paired = TRUE, conf.level = 0.95)
```

```
##
##  Paired t-test
##
## data:  zenski$student and zenski$nagrada_student
## t = -0.33591, df = 21, p-value = 0.7403
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -48.04850  34.68487
## sample estimates:
## mean of the differences
##               -6.681818
```

Iz ovih testova vidimo da nema prevelike razlike u predikcijama između muškaraca i žena.

```
t.test(muski$MI, zenski$MI, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  muski$MI and zenski$MI
## t = 2.6463, df = 44.702, p-value = 0.0112
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6102206 4.5016675
## sample estimates:
## mean of x mean of y
##  18.94231  16.38636
```

Iz ovog testa vidimo da postoji razlika u rezultatu na međuispitu između muškaraca i žena.

4.2 ANOVA

```
aov <- manova(cbind(student,cijela_grupa,samo_prijatelji)~nagrada,data=podaciNagrada)
summary(aov)
```

```
##            Df  Pillai approx F num Df den Df Pr(>F)
## nagrada     1 0.01827   1.2159      3    196 0.3051
## Residuals 198
```

```
print("-------------------------------------------------------------------------------------")
```

```
## [1] "-------------------------------------------------------------------------------------"
```

```
aov <- aov(MI~SPOL,data=podaci)
summary(aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## SPOL         1  112.1  112.10   5.045 0.0269 *
## Residuals   98 2177.5   22.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print("-------------------------------------------------------------------------------------")
```

```
## [1] "-------------------------------------------------------------------------------------"
```

```
aov <- manova(cbind(student,cijela_grupa,samo_prijatelji,
                    nagrada_student,nagrada_cijela_grupa,
                    nagrada_samo_prijatelji)~SPOL,data=podaci)
summary(aov)
```

```
##           Df   Pillai approx F num Df den Df Pr(>F)
## SPOL       1 0.055749  0.91513      6     93 0.4877
## Residuals 98
```

Proveli smo 3 ANOVA testa. U prvom testu smo testirali ovisi li nagrada o predikcijama studenta. Vrijednost statistike je bila 0.30, što znači da ne možemo odbaciti nul hipotezu da predikcije ne ovise o nagradi.

U drugom testu smo testirali ovisnost rezultata na međuispitu o spolu. Kao što smo prije vidjeli u deskriptivnoj statistici da muškarci imaju bolje rezultate na međuispitu od žena, ANOVA nam je potvrdila da možemo odbaciti pretpostavku da rezultati na međuispitu na ovise o spolu.

U trećem testu smo vidjeli da predikcije ne ovise o spolu.

5 STROJNO UČENJE

Pregled strojnog učenja napravljen je prema (Alpaydin 2010).

5.1 Linearna regresija

Iz deskriptivna analize smo vidjeli da MI nema skoro nikakve korelacije s predikacijama, dok sve ostale varijable imaju veliku pozitivnu korelaciju. Iz ovoga možemo pretpostaviti da ćemo dobiti nikakvu pametnu linearnu regresiju za međuispit, dok za ostale varijable linearna regresija će vjerojatno uključivati sve varijable.

Linearnu regresiju ćemo napraviti uz pomoć iterativne selekciju u dva smjera. Prvo ćemo proglasiti da varijabla ovisi o svim varijablama, a onda u drugom slučaju da varijabla ne ovisi o ničemu. U prvom slučaju ćemo odbacivati varijable jednu po jednu pa gledati kad je najbolja R squared mjera, a u drugom slučaju dodavati varijable jednu po jednu. Ovo ćemo provesti za svih 7 varijabli.

```
##

lm_sve <- lm(MI ~ ., data=podaciLinear)
lm_prazan <- lm(MI ~ 1, data=podaciLinear)

lm1 <- stepAIC(lm_sve, direction="backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan),
               direction="forward", trace = 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = MI ~ 1, data = podaciLinear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.880  -3.380   0.370   3.745   9.620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3800     0.4809   38.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.809 on 99 degrees of freedom
```

```
#summary(lm2)

###
lm_sve <- lm(student ~ ., data=podaciLinear)
lm_prazan <- lm(student ~ 1, data=podaciLinear)

lm1 <- stepAIC(lm_sve, direction="backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan),
               direction="forward", trace = 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = student ~ cijela_grupa + samo_prijatelji + nagrada_student +
```

```
##     nagrada_samo_prijatelji, data = podaciLinear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -236.75  -57.46   -1.49   30.32  463.74
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7.94709   45.00664   0.177    0.860
## cijela_grupa              0.10236    0.07009   1.460    0.147
## samo_prijatelji           0.39370    0.08589   4.584 1.39e-05 ***
## nagrada_student           0.90637    0.06919  13.100  < 2e-16 ***
## nagrada_samo_prijatelji -0.40578    0.09423  -4.306 4.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108.1 on 95 degrees of freedom
## Multiple R-squared:  0.7853, Adjusted R-squared:  0.7763
## F-statistic: 86.88 on 4 and 95 DF,  p-value: < 2.2e-16
```

```r
#summary(lm2)

###
lm_sve <- lm(cijela_grupa ~ ., data=podaciLinear)
lm_prazan <- lm(cijela_grupa ~ 1, data=podaciLinear)

lm1 <- stepAIC(lm_sve, direction="backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan),
               direction="forward", trace = 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = cijela_grupa ~ student + samo_prijatelji + nagrada_cijela_grupa +
##     nagrada_samo_prijatelji, data = podaciLinear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -627.75  -79.61    3.80   73.91  376.53
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             197.14078   52.67479   3.743 0.000312 ***
## student                   0.12769    0.07693   1.660 0.100223
## samo_prijatelji           0.28208    0.11085   2.545 0.012545 *
## nagrada_cijela_grupa      0.59688    0.10779   5.537 2.72e-07 ***
## nagrada_samo_prijatelji  -0.37988    0.13044  -2.912 0.004470 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 136.2 on 95 degrees of freedom
## Multiple R-squared:  0.4727, Adjusted R-squared:  0.4505
## F-statistic: 21.29 on 4 and 95 DF,  p-value: 1.472e-12
```

```r
#summary(lm2)

###
lm_sve <- lm(samo_prijatelji ~ ., data=podaciLinear)
lm_prazan <- lm(samo_prijatelji ~ 1, data=podaciLinear)

lm1 <- stepAIC(lm_sve, direction="backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan),
               direction="forward", trace = 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = samo_prijatelji ~ student + cijela_grupa + nagrada_student +
##     nagrada_samo_prijatelji, data = podaciLinear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -273.80  -67.22   -7.16   41.97  429.94
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              20.14253   48.61331   0.414  0.67956
## student                   0.46000    0.10036   4.584 1.39e-05 ***
## cijela_grupa              0.15691    0.07490   2.095  0.03884 *
## nagrada_student          -0.38469    0.11891  -3.235  0.00167 **
## nagrada_samo_prijatelji   0.73806    0.08165   9.040 1.87e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 116.9 on 95 degrees of freedom
## Multiple R-squared:  0.6783, Adjusted R-squared:  0.6647
## F-statistic: 50.07 on 4 and 95 DF,  p-value: < 2.2e-16
```

```r
#summary(lm2)

###

###
lm_sve <- lm(nagrada_student ~ ., data=podaciLinear)
lm_prazan <- lm(nagrada_student ~ 1, data=podaciLinear)

lm1 <- stepAIC(lm_sve, direction="backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan),
               direction="forward", trace = 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = nagrada_student ~ student + samo_prijatelji + nagrada_cijela_grupa +
##     nagrada_samo_prijatelji, data = podaciLinear)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -228.433  -56.436   -6.825   38.641  241.213
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              23.38362   36.09300   0.648  0.51863
## student                   0.68139    0.05271  12.927  < 2e-16 ***
## samo_prijatelji          -0.25043    0.07595  -3.297  0.00137 **
## nagrada_cijela_grupa      0.16628    0.07386   2.251  0.02667 *
## nagrada_samo_prijatelji  0.37394    0.08938   4.184 6.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.29 on 95 degrees of freedom
## Multiple R-squared:  0.8106, Adjusted R-squared:  0.8027
## F-statistic: 101.7 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
#summary(lm2)

###
lm_sve <- lm(nagrada_cijela_grupa ~ ., data=podaciLinear)
lm_prazan <- lm(nagrada_cijela_grupa ~ 1, data=podaciLinear)

lm1 <- stepAIC(lm_sve, direction="backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan),
               direction="forward", trace = 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = nagrada_cijela_grupa ~ student + cijela_grupa +
##     nagrada_student + nagrada_samo_prijatelji, data = podaciLinear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -267.98  -57.77   -5.18   32.67  370.31
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -12.89926   45.52605  -0.283  0.77753
## student                   -0.13328    0.09398  -1.418  0.15942
## cijela_grupa               0.39544    0.07014   5.638 1.76e-07 ***
## nagrada_student            0.30877    0.11136   2.773  0.00669 **
## nagrada_samo_prijatelji   0.48047    0.07646   6.284 9.94e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109.5 on 95 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6913
## F-statistic: 56.42 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
#summary(lm2)
###
lm_sve <- lm(nagrada_samo_prijatelji ~ ., data=podaciLinear)
lm_prazan <- lm(nagrada_samo_prijatelji ~ 1, data=podaciLinear)
```

```r
lm1 <- stepAIC(lm_sve, direction="backward", trace = 0)
lm2 <- stepAIC(lm_prazan, scope = list(upper = lm_sve, lower = lm_prazan),
               direction="forward", trace = 0)
summary(lm1)
```

```
##
## Call:
## lm(formula = nagrada_samo_prijatelji ~ student + cijela_grupa +
##     samo_prijatelji + nagrada_student + nagrada_cijela_grupa,
##     data = podaciLinear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -288.47  -44.98    5.56   49.85  419.73
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          81.13061   39.21219   2.069  0.04129 *
## student              -0.27520    0.08738  -3.150  0.00219 **
## cijela_grupa         -0.16251    0.07097  -2.290  0.02427 *
## samo_prijatelji       0.51790    0.06569   7.884 5.58e-12 ***
## nagrada_student       0.37030    0.09933   3.728  0.00033 ***
## nagrada_cijela_grupa  0.40077    0.08040   4.984 2.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 96.28 on 94 degrees of freedom
## Multiple R-squared:  0.7563, Adjusted R-squared:  0.7434
## F-statistic: 58.35 on 5 and 94 DF,  p-value: < 2.2e-16
```

```r
#summary(lm2)
```

Sve ono što smo pretpostavili prije linearne regresije se pokazalo istinito. Također, regresija unaprijed i unazad nam je dala jednake rezultate pa je nismo ispisivali radi preglednosti ispisa.

5.2 Logistička regresija

Logistička regresija je metoda za klasifikaciju podataka u diskretne klase. Logička regresija je takva da predviđa vjerojatno ishoda kada postoje samo dvije mogućnosti (istina ili laž). Ako želimo logičku regresiju iskoristiti za predviđanje n mogućnosti, problem ćemo razložiti u n problema. Svaki problem je zasebna klasifikacija u dvije mogućnosti, odnosno dali podatak spada u tu klasu ili ne spada. Za onu klasu za koju dobijemo najveću vjerojatnost da podatak pripada u nju, proglasit ćemo da podatak pripada toj klasi.

Varijabla koju želimo predvidjeti logističkom regresijom mora biti nominalna, dok varijable o kojima želimo da model zavisi moraju biti numeričke.

S logističkom regresijom ćemo pokušati predvidjeti spol u ovisnosti o rezultatu na međuispitu, spol u ovisnosti o svemu te nagradu o ovisnosti o predikcijama.

```
logisticReg <- multinom(SPOL ~ MI, data =podaci,family=binomial)
```

```
## # weights:  3 (2 variable)
## initial  value 69.314718
## final  value 50.271244
## converged
```

```
summary(logisticReg)
```

```
## Call:
## multinom(formula = SPOL ~ MI, data = podaci, family = binomial)
##
## Coefficients:
##                 Values  Std. Err.
## (Intercept) -0.7023369 0.91604927
## MI           0.1112004 0.05166162
##
## Residual Deviance: 100.5425
## AIC: 104.5425
```

```
pred <- predict(logisticReg,podaci)

podaci %>% dplyr::select(SPOL) -> results
results$pred <- pred

error_2 <- results$SPOL==results$pred

summary(error_2)
```

```
##    Mode   FALSE    TRUE
## logical      22      78
```

```
logisticReg <- multinom(SPOL ~ ., data =podaci,family=binomial)
```

```
## # weights:  9 (8 variable)
## initial  value 69.314718
## iter  10 value 47.393036
## final  value 47.344938
## converged
```

```
summary(logisticReg)
```

```
## Call:
## multinom(formula = SPOL ~ ., data = podaci, family = binomial)
##
```

```
## Coefficients:
##                                Values    Std. Err.
## (Intercept)              -2.2191929891 0.0009973494
## student                   0.0031558482 0.0031798659
## cijela_grupa              0.0009758084 0.0022653603
## samo_prijatelji          -0.0022166696 0.0023827034
## nagrada_student          -0.0010204439 0.0034023449
## nagrada_cijela_grupa     -0.0034512958 0.0025776328
## nagrada_samo_prijatelji   0.0050972231 0.0029881378
## MI                        0.1149703186 0.0415308030
##
## Residual Deviance: 94.68988
## AIC: 110.6899
```

```
pred <- predict(logisticReg,podaci)

podaci %>% dplyr::select(SPOL) -> results
results$pred <- pred

error_2 <- results$SPOL==results$pred

summary(error_2)
```

```
##    Mode   FALSE    TRUE
## logical      24      76
```

```
podaciNagrada %>% dplyr::select(-SPOL,-MI) -> podaciNagradaBezSpola
podaciNagradaBezSpola$nagrada <- ifelse(podaciNagradaBezSpola$nagrada,"NAGRADA",
                                        "NEMA NAGRADE")
podaciNagradaBezSpola$nagrada <- as.factor(podaciNagradaBezSpola$nagrada)

logisticReg <- multinom(nagrada ~ ., data =podaciNagradaBezSpola,family=binomial)
```

```
## # weights:  5 (4 variable)
## initial  value 138.629436
## final  value 136.762387
## converged
```

```
summary(logisticReg)
```

```
## Call:
## multinom(formula = nagrada ~ ., data = podaciNagradaBezSpola,
##     family = binomial)
##
## Coefficients:
##                         Values    Std. Err.
## (Intercept)       0.3452965592 0.5468121456
## student           0.0005574171 0.0008713065
## cijela_grupa     -0.0019683580 0.0010504699
## samo_prijatelji   0.0008180253 0.0010349294
##
## Residual Deviance: 273.5248
## AIC: 281.5248
```

```
pred <- predict(logisticReg,podaciNagradaBezSpola)

podaciNagradaBezSpola %>% dplyr::select(nagrada) -> results
```

```r
results$pred <- pred

error_2 <- results$nagrada==results$pred

summary(error_2)
```

```
##    Mode   FALSE    TRUE
## logical      93     107
```

S predikcijom o spolu nismo ništa pametno dobili. Za predikcije o spolu smo dobili jednak rezultat kao da smo proglasili sve predikcije da su predikcije muškaraca, a za predikciju za nagradu smo dobili malo bolji rezultat nego da smo proglasili da su sve s nagradom. Zanimljivo je i promotriti koeficijente logističke regresije. Možemo vidjeti da ovisnost rezultata međuispita o spolu ima pozitivan koeficijent, te da predikcija nagrada najviše ovisi o predikciji za cijelu grupu s negativnim koeficijentom logističke regresije.

5.3 Stroj potpornih vektora

Stroj potpornih vektora je jedan on najpopularnijih modela strojnih učenja koji se danas koristi pri bilo kojoj klasifikaciji. Stroj potpornih vektora je klasifikator koji konstrukcijom hiperravnine u visoko-dimenzionalnom prostoru stvara model koji predviđa kojoj klasi pripada novi uzorak. Stroj potpornih vektora na ulaz dobiva podatke povezane s klasom kojoj pripadaju te ih prikazuje kao točke u prostoru raspoređene način da su točke koje predstavljaju podatke koji pripadaju različitim klasama međusobno što razmaknutije.

Zadatak stroja potpornog vektora je odabrati optimalnu hiperravninu razdvajanja. Optimalna hiperravnina razdvajanja je ona koja ostavlja najviše slobodnog prostora između klasa, tj. maksimizira marginu između hiperravnine i klasa. Model koristi vektore podataka za određivanje maksimalne margine i ti vektori se nazivaju potporni vektori.

```
svm <- svm(SPOL ~ MI, data =podaci)

pred <- predict(svm,podaci)

podaci %>% dplyr::select(SPOL) -> results
results$pred <- pred

error_2 <- results$SPOL==results$pred

summary(error_2)
```

```
##    Mode   FALSE    TRUE
## logical      22      78
```

```
svm <- svm(SPOL ~ ., data =podaci)

pred <- predict(svm,podaci)

podaci %>% dplyr::select(SPOL) -> results
results$pred <- pred

error_2 <- results$SPOL==results$pred

summary(error_2)
```

```
##    Mode   FALSE    TRUE
## logical      22      78
```

```
podaciNagrada %>% dplyr::select(-SPOL,-MI) -> podaciNagradaBezSpola
podaciNagradaBezSpola$nagrada <- ifelse(podaciNagradaBezSpola$nagrada,"NAGRADA",
                                        "NEMA NAGRADE")
podaciNagradaBezSpola$nagrada <- as.factor(podaciNagradaBezSpola$nagrada)

svm <- svm(nagrada ~ ., data = podaciNagradaBezSpola)

pred <- predict(svm,podaciNagradaBezSpola)

podaciNagradaBezSpola %>% dplyr::select(nagrada) -> results
results$pred <- pred

error_2 <- results$nagrada==results$pred

summary(error_2)
```

```
##    Mode   FALSE    TRUE
## logical      79     121
```

Opet iz predikcije o spolu nismo dobili ništa bolje nego da smo proglasili da su svi podaci predikcije muškaraca. Kod predikcije o nagradi uz pomoć SVM-a smo dobili prilično solidan rezultat te smo uz točnost od 60.5% uspjeli predvidjeti jesu li predikcije o broju stranica knjige donesene ako je u pitanju nagrada ili nije.

6 ZAKLJUČAK

Analiza ljudskog ponašanja je uvijek vrlo kontroverzna stvar. Još je kontroverzije kad promatramo primalne nagone u ljudi kao što su nagrada, moć, utjecaj i sl. Glavna motivacija ovog rada je bila vidjeti postoje li razlike u procjenama studenata FER-a o broju stranica knjige ako smo su dobili nagradu ili ne.

Iz statističke analize nismo uspjeli dokazati da postoje razlike u procjenama vezano za nagrade. Postoje neke sitne razlike u procjenama, ali ništa da bi sa sigurnošću mogli reći da nagrada utječe na procjenu studenata FER-a

U budućnosti, ovaj rad se sigurno može i treba revidirati. Prvo, veći dataset je nužan. 100 studenata je jako mali uzorak da bi se sa sigurnošću nešto moglo dokazati. Također, studente se treba uvjeriti da je istraživanje ozbiljno i da ne unose procjene skroz nasumično.

LITERATURA

Alpaydin, Ethem. 2010. *Introduction to Machine Learning.*

Diez, M. David, D. Christopher Barr, and Mine Cetinkaya-Rundel. 2015. *OpenIntro Statistics.*

Zimbardo, P.G. 1971. "The Power and Pathology of Imprisonment."