1. zadaća – Inteligentni sustavi 2

Radila: Nola Čumlievski

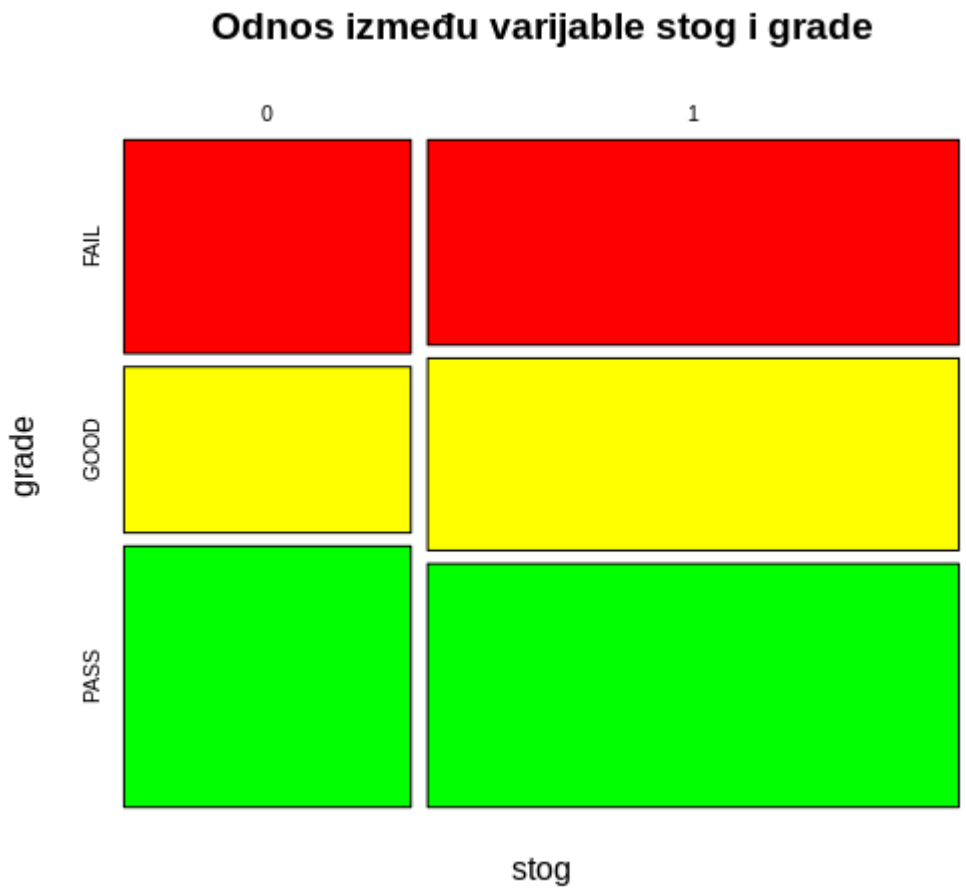
a) Utjecaj varijable „stog” na ocjenu studenta

```
attach(edukacija90)
podaci=data.frame(stog,grade)
podaci
boja=c("red", "yellow","green")
mosaicplot(table(podaci),color=boja, main="Odnos između varijable stog i grade")
```

Slika 1: Utjecaj varijable "stog" na varijablu "grade" - kod

Na slici 1. prikazan je kod za vizualizaciju utjecaja varijable „*stog*” na konačnu ocjenu studenta. U ovoj zadaći prvenstveno se analizira koliko pohađanje prezentacije Stog utječe na konačnu ocjenu studenta, a zatim i utjecaj iste varijable na ocjenu u kombinaciji s drugim varijablama (*labs, videos, lectures*,...). Prvi korak bio je download baze „*edukacija90*” te import iste u Rstudio (File > Import Dataset > From text (base)). Kada se obavi import potrebno je koristiti funkciju „*attach”* da bi se služili imenima varijabli bez potrebe da naglašavamo u kojoj se bazi varijabla nalazi (bez *attach* fukcije umjesto „stog” koristili bi izraz „edukacija90$stog” za pristup varijabli unutar određene baze).

Slijedeći korak je izrada data frame-a s potrebnim podacima („*stog”* i „*grade”*). Vektor „*boja”* napravljen je da bi svaka ocjena u grafikonu imala različitu boju (crvena kao fail, žuta kao good i zelena kao pass). Grafikon je izrađen pomoću funkcije „*mosaicplot”* koja kao argumente ima data frame u obliku tablice, argument „*color”* koji je jednak vektoru koji je prethodno iniciran i argument „*main”* za naslov grafikona. Iz vizualizacije vidljivo je da varijabla *„stog”* nema velik utjecaj na ocjenu studenata, s obzirom da je otprilike jednak broj studenata, odslušali prezentaciju ili ne, palo i prošlo kolegij. Grafikon je prikazan na slici 2.



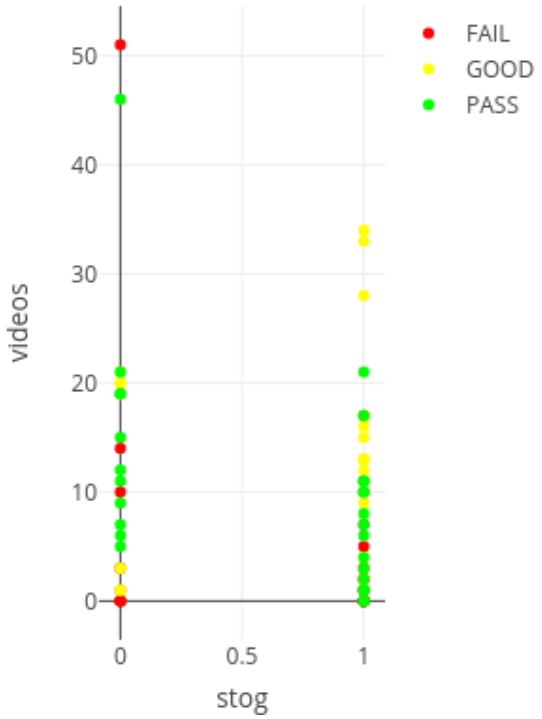
Slika 2: Odnos između varijable stog i grade - grafikon

b) Utjecaj varijable „*stog*” u kombinaciji s drugim varijablama

```
#scatter plotovi sa više utjecajnih varijabli
#utjecaja varijabli videos i stog na konačnu ocjenu studenta
fig=plot_ly(data=edukacija90, x=~stog, y=~videos, color=~grade, colors=boja)
fig
#utjecaj varijable labs i stog na ocjenu studenta
fig=plot_ly(data=edukacija90, x=~stog, y=~labs, color=~grade, colors=boja)
fig
#utjecaj varijabli selfassesm i stog na ocjenu studenta
fig=plot_ly(data=edukacija90, x=~stog, y=~selfassesm, color=~grade, colors=boja)
fig
#utjecaj varijabli stog i lectures na ocjenu studenta
fig=plot_ly(data=edukacija90, x=~stog, y=~lectures, color=~grade, colors=boja)
fig
#utjecaj varijabli stog i quizzes na ocjenu studenta
fig=plot_ly(data=edukacija90, x=~stog, y=~quizzes, color=~grade, colors=boja)
fig
```

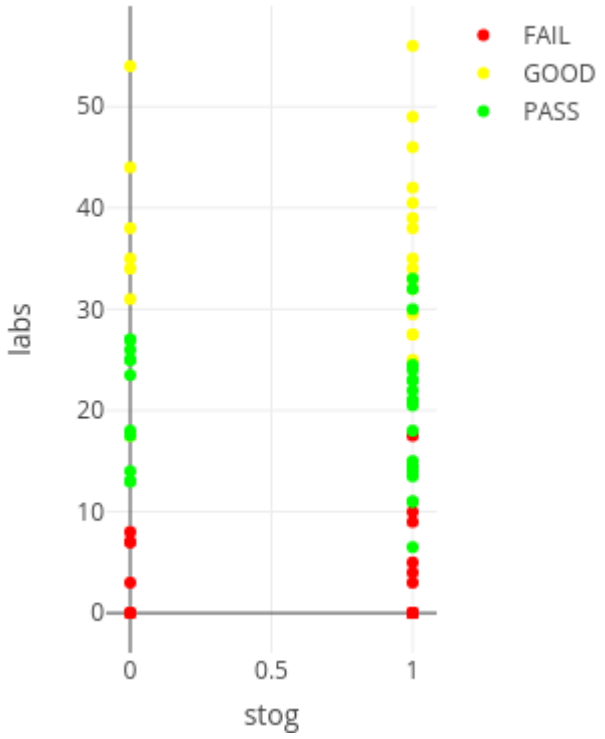
Slika 3: Utjecaj varijable stog u kombinaciji s drugim varijablama - kod

Na slici 3. vidljiv je kod korišten za vizualizaciju dijagrama raspršenja. Za izradu dijagrama korištena je funkcija *„plot_ly”* iz istoimene biblioteke. Funkcija ima slijedeće argumente: *„data”* kojim označavamo iz koje baze prikupljamo podatke za vizualizaciju, *„x” i „y”* kojim označavamo koje varijable želimo prikazane na x i y osi, „*color”* kojim označavamo varijablu za koju želimo vizualizirati utjecaj prethodne dvije varijable x i y, gleda se podjela prema njenim vrijednostima – varijabla klasa (u legendi i na grafikonu prikazana je posebna boja za svaku od vrijednosti te varijable) te argument „*colors”* kojim označavamo koje boje želimo koristiti u vizualizaciji (ovdje je korišten vektor „*boja”* koji je iniciran u prethodnom bloku koda).



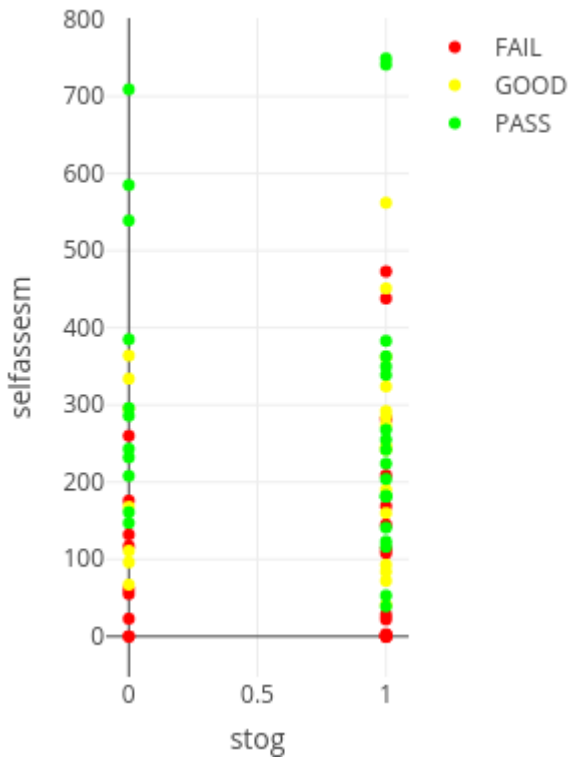
Slika 4: Utjecaj varijabli stog i videos na ocjenu - grafikon

Na slici 4. prikazan je grafikon (scatter plot) koji nam prikazuje utjecaj varijable *„stog" i „videos"* na konačnu ocjenu studenata. Iz priloženog vidljivo je da je velika količina studenata koji su odlušali prezentaciju Stog i koji su pokretali snimke predavanja (otprilike 0-35 puta) imalo ocjenu *„good" i „pass",* tj. prošli su kolegij, dok je 3 studenta koja nisu odslušali prezentaciju stog i pogledali su snimke predavanja do 15 puta palo kolegij. Primjer točke gdje je student/ica prošao/la kolegij – (1, 21) gdje jedinica označava odslušanu pretentaciju, a 21 broj pokretanja snimki predavanja. Primjer gdje student/ica nije prošao/la kolegij – (1, 5).



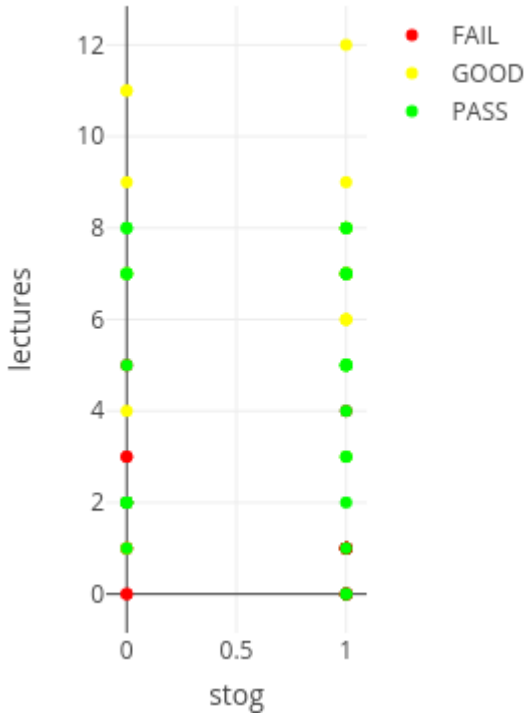
Slika 5: Utjecaj varijabli stog i labs na ocjenu studenta - grafikon

Na slici 5. prikazan je grafikon koji nam prikazuje utjecaj varijable *„stog" i „labs"* na konačnu ocjenu studenata. Iz priloženog vidljivo je da na konačnu ocjenu studenta dosta više utjecalo to koliko je određeni student ostvario bodova na vježbama nego da li je odslušao prezentaciju Stog, dakle, u kombinaciji sa varijablom *„labs",* varijabla *„stog"* nema velik utjecaj na konačnu ocjenu studenta (čak je veći broj studenata koji su pali kolegij, a odslušali su prezentaciju). Također je velik broj studenata koji su odslušali prezentaciju Stog i ostvarili više od 10 bodova na vježbama, a da su prošli kolegij. Primjer točke gdje je student/ica prošao/la kolegij – (0, 27). Primjer gdje student/ica nije prošao/la kolegij – (1, 17.5).



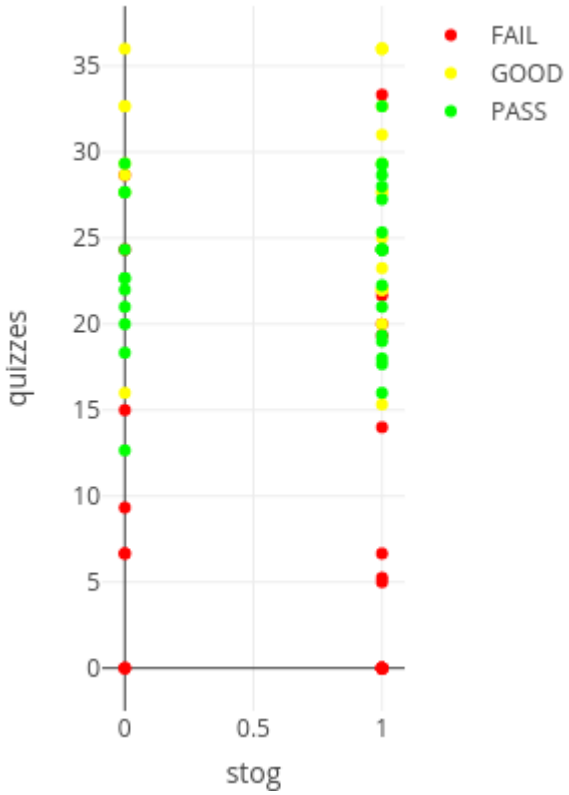
Slika 6: Utjecaj varijabli stog i selfassesm na ocjenu studenta - grafikon

Na slici 6. prikazan je grafikon koji nam prikazuje utjecaj varijable *„stog" i „selfsassesm"* na konačnu ocjenu studenata, gdje *„selfsassesm"* predstavlja mjeru aktivnosti (broj klikova u samoprovjeri). Iz grafikona vidljivo je, u odnosu na varijablu *„stog",* da je broj studenata koji su pali kolegij otprilike jednak, dok su većina studenata koji su prošli kolegij imali srednje do visok broj klikova u okviru samoprovjere (uz dvije iznimke – (1, 438 i 1, 473)). Zaključak: u kombinaciji sa varijablom *„selfassesm",* varijabla *„stog"* nema velik utjecaj na konačnu ocjenu studenta. Primjer točke gdje je student/ica prošao/la kolegij – (0, 585). Primjer gdje student/ica nije prošao/la kolegij – (1, 209).



Slika 7: Utjecaj varijable stog i lectures na ocjenu studenta - grafikon

Na slici 7. prikazan je grafikon koji nam prikazuje utjecaj varijable *„stog" i „lectures"* na konačnu ocjenu studenata, gdje *„lectures"* označava broj bodova koji su studenti dobili u okviru aktivnosti. Iz grafikona je vidljivo da u kombinaciji s varijablom *„lectures",* varijabla *stog"* ima velik utjecaj na konačnu ocjenu studenta, s obzirom da dva studenta koja su pala kolegij nisu odslušali prezentaciju Stog i imali su vrlo nizak broj bodova u okviru aktivnosti. Primjer točke gdje je student/ica prošao/la kolegij – (1, 6). Primjer gdje student/ica nije prošao/la kolegij – (0, 3).



Slika 8: Utjecaj varijable stog i quizzes na ocjenu - grafikon

Na slici 8. prikazan je grafikon koji nam prikazuje utjecaj varijable *„stog" i „quizzes"* na konačnu ocjenu studenata, gdje varijabla *„quizzes"* označava broj bodova koje je student ostvario na oba kviza. Iz grafikona je vidljivo da je velika većina studenata koja je prošla kolegij imala veći broj ostvarenih bodova u kvizovima, dok varijabla *„stog"* nema velik utjecaj na prolaz studenta/ice (skoro je jednak broj studenata koji su pali kolegij, bez obzira jesu ili nisu odslušali prezentaciju Stog). Zaključak: u kombinaciji sa varijablom *„quizzes",* varijabla *„stog"* nema velik utjecaj na konačnu ocjenu studenta. Primjer točke gdje je student/ica prošao/la kolegij – (0, 22). Primjer gdje student/ica nije prošao/la kolegij – (1, 21.66).

c) Logistička regresija na skupu „edukacija88.csv“

```
#logistička regresija
attach(edukacija88)
#priprema podataka za izradu modela
train=id<60
train
test=!train
treniranje=edukacija88[train,]
testiranje=edukacija88[test,]
testiranje
y=grade[test]
y
```

Slika 9: Priprema podataka za izradu modela LR

Na slici 9. vidljiv je kod za pripremu podataka kako bi mogli izraditi modele logističke regresije. Prvi korak je import i attach baze *„edukacija88”*. Zatim slijedi označavanje indeksa kako bi bazu podijelili na skup za treniranje i skup za testiranje. Za treniranje sam odabrala prvih 60 redova baze (ne uključujući 60. red), a ostali služe za testiranje. Slijedi dohvaćanje opservacije u bazi *„edukacija88”* koje se zatim svrstavaju u navedene skupove. Također je potrebno inicirati skup vrijednosti nad kojim će se vršiti predikcija (*y=grade[test]*). S obzirom da u bazi ima 77 opservacija, kada pokrenemo *„y”* dobit ćemo 18 vrijednosti levela *„PASS”* ili *„FAIL”*.

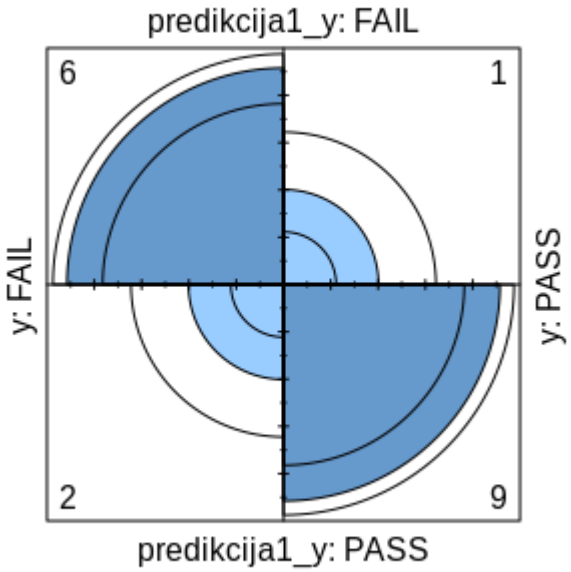
1. model

```
#model1 - uzima u obzir varijable stog i lectures za predikciju ocjene
library(ISLR)
model1=glm(formula=grade~lectures+stog, family="binomial", data=treniranje)
predikcija=predict(model1,testiranje,type="response")
predikcija1_y=rep("FAIL",length(ytest))
predikcija1_y
predikcija1_y[predikcija>0.5]="PASS"
predikcija1_y
prva=table(predikcija1_y,y)
fourfoldplot(prva)
mean(predikcija1_y!=y)
```

Slika 10: Prvi model - kod

Na slici 10. prikazan je kod za izradu prvog modela logističke regresije koji uzima u obzir varijable *„stog” i „lectures”* za predikciju ocjene studenta. Za izradu modela korištena je funkcija *„glm”* koja se koristi za izradu generaliziranih linearnih modela. Navedena funckija sadrži slijedeće parametre: *„formula”* u kojem označavamo koju vrijednost želimo predvidjeti i varijable koje želimo uzeti u obzir pri predikciji, *„family”* specificira porodicu modela (ako se ne odredi porodica izvest će se obična linearna regresija) te *„data”* u kojem označavamo skup za

treniranje, tj. skup na kojem će se model trenirati. Za predikciju koristimo funkciju *„predict”* u kojoj navodimo naziv modela, skup za testiranje i tip (za glm modele koristi se tip *„response”*). Kako funkcija vraća numeričke vrijednosti (predviđenu vrijednost), stvaramo vektor *„predikcija1_y”* koji ponavlja vrijednost *„FAIL”* kolika je dužina varijable *„y”* te se zatim ažurira na način da se upiše vrijednost *„PASS”* na pozicije za koje je predviđena vjerojatnost veća od 0.5. Varijabla *„prva”* predstavlja matricu predviđenih vrijednosti za y i stvarnih vrijednosti y. Za vizualizaciju matrice može se koristiti ugrađena funkcija *„fourfoldplot”,* a funkcija *„mean”* koristi se za izračun postotka greške modela (1.666667 ili ~16%). Postupak je ponovljen za još 4 modela (ukupno 5 modela), navodeći drugačije prediktore kod funckije *„glm”*. Na slici 11. prikazana je matrica konfuzije za prvi model.



Slika 11: Prvi model - matrica konfuzije

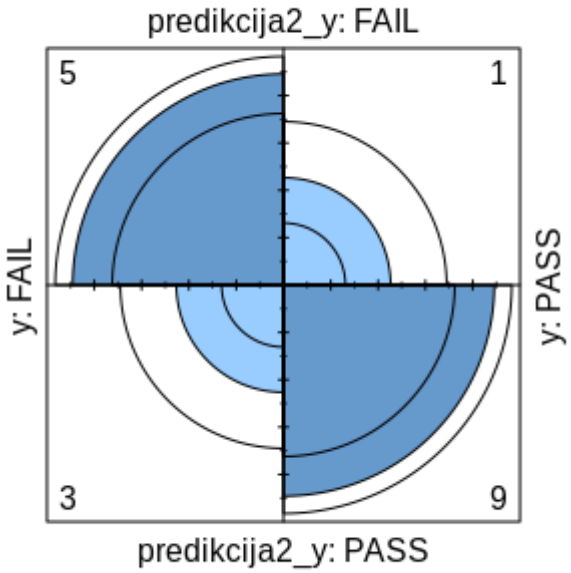
Dijagonala matrice označava broj točno predviđenih vrijednosti. Prvi model točno je predvidio 15 vrijednosti od 18. Kao što je prikazano, matrica se sastoji od 4 kvadranta: prvi kvadrant označava stvarnu i točno predviđenu klasu *„FAIL”,* drugi označava stvarnu klasu *„PASS”* koja je predviđena pogrešno kao klasa *„FAIL”,* treći kvadrant označava stvarnu klasu *„FAIL”* koja je netočno predviđena kao *„PASS”* i četvrti kvadrant koji označava stvarnu i točno predviđenu klasu *„PASS”.*

2. model

U drugom modelu korišteni su prediktori *„stog” i „selfassesm”*. Postotak pogreške drugog modela iznosi 0.2777778 (~27%), što je znatno veća pogreška od prvog modela.

```r
#model2  - uzima u obzir varijable selfassesm i stog za predikciju ocjene
model2=glm(formula=grade~selfassesm+stog, family="binomial", data=treniranje)
predikcija2=predict(model2,testiranje,type="response")
predikcija2_y=rep("FAIL",length(ytest))
predikcija2_y[predikcija2>0.5]="PASS"
predikcija2_y
druga=table(predikcija2_y,y)
fourfoldplot(druga)
mean(predy2!=ytest)
```

Slika 12: Drugi model - kod



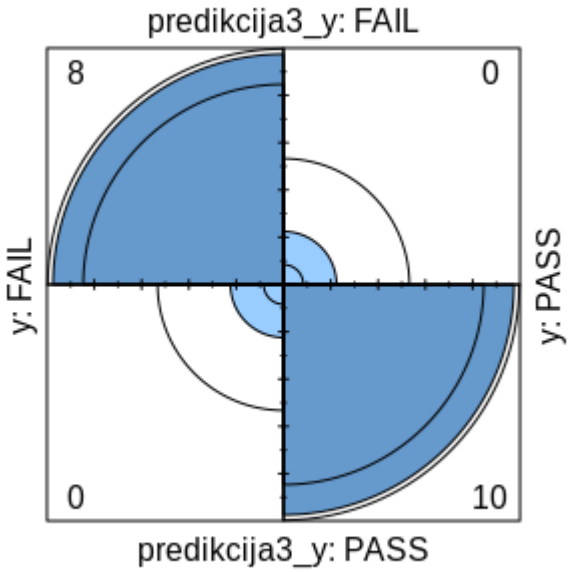
Sa slike 13. vidljive su predikcije drugog modela. 14 observacija je točno predviđeno, dok je 4 opservacije model predvidio netočno.

3. model

U trećem modelu korišteni su prediktori *„stog", „selfassesm", „labs", „lectures" i „quizzes"*. Postotak pogreške trećeg modela iznosi 0 (0%), što znači da je navođenjem navedenih 5 varijabli kao prediktora dobiven najbolji rezultat modela za bazu podataka *„edukacija88"*.

```r
#model3 - uzima u obzir varijable selfassesm, stog, labs, lectures i quizzes za predikciju ocjene
model3=glm(formula=grade~selfassesm+stog+labs+lectures+quizzes, family="binomial", data=treniranje)
predikcija3=predict(model3,testiranje,type="response")
predikcija3_y=rep("FAIL",length(ytest))
predikcija3_y[predikcija3>0.5]="PASS"
predikcija3_y
tri=table(predikcija3_y,y)
fourfoldplot(tri)
mean(predy3!=ytest)
```

Slika 14: Treći model - kod



Slika 15: Treći model - matrica konfuzije

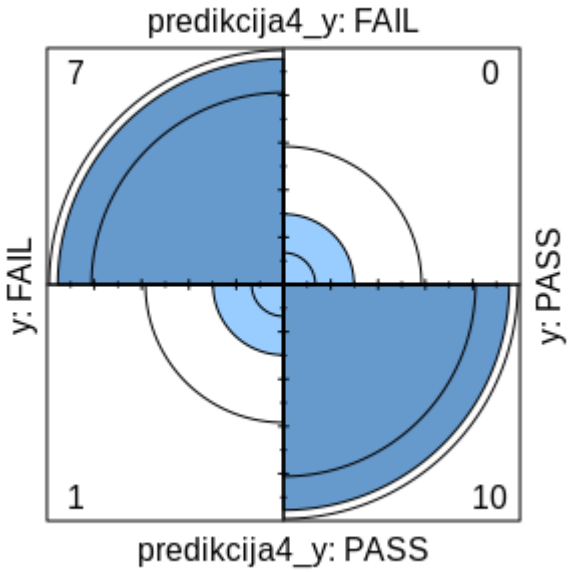
Iz priložene matrice vidimo da je broj točnih opservacija 18/18.

4. model

U četvrtom modelu korišteni su prediktori *„stog" i „labs*. Postotak pogreške četvrtog modela iznosi 0.05555556 (~5%).

```
#model4 - uuzima u obzir varijable stog i labs za predikciju ocjene
model4=glm(formula=grade~stog+labs, family="binomial", data=treniranje)
predikcija4=predict(model4,testiranje,type="response")
predikcija4_y=rep("FAIL",length(ytest))
predikcija4_y[predikcija4>0.5]="PASS"
predikcija4_y
cetvrta=table(predikcija4_y,y)
fourfoldplot(cetvrta)
mean(predy4!=ytest)
```

Slika 16: Četvrti model - kod



Slika 17: Četvrti model - matrica konfuzije

Iz priložene matrice vidimo da je model netočno predvidio jednu „*FAIL”* vrijednost.

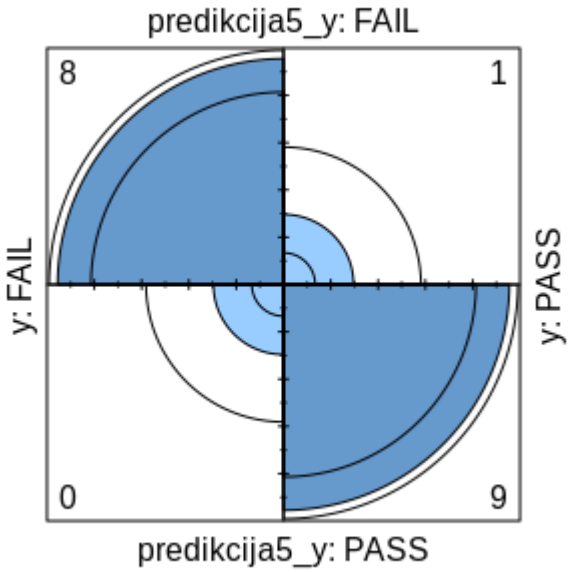
5. model

U petom modelu korišteni su prediktori iz trećeg modela + prediktori *„forum” i „demons”*.

Postotak pogreške petog modela također iznosi 0.05555556 (~5%).

```
#model5 - uzima u obzir varijable stog, labs, videos, lectures, selfassesm, forum i demons za predikciju ocjene
model5=glm(formula=grade~stog+labs+videos+lectures+selfassesm+forum+demons, family="binomial", data=treniranje)
predikcija5=predict(model5,testiranje,type="response")
predikcija5_y=rep("FAIL",length(ytest))
predikcija5_y[predikcija5>0.5]="PASS"
predikcija5_y
peta=table(predikcija5_y,y)
mean(predy5!=ytest)
fourfoldplot(peta)
```

Slika 18: Peti model - kod

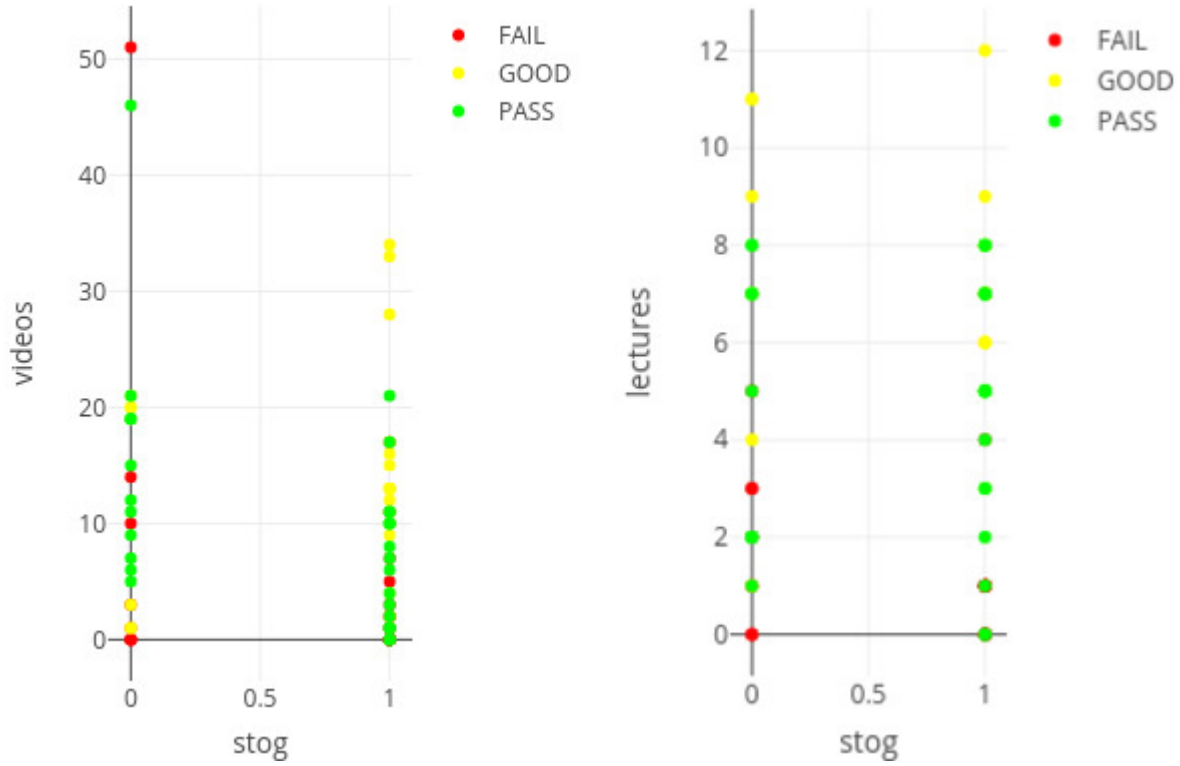


Slika 19: Peti model - matrica konfuzije

Za raliku od četvrtog modela, peti model predvidio je netočno klasu *„PASS”,* no postotak pogreške je identičan (~5%).

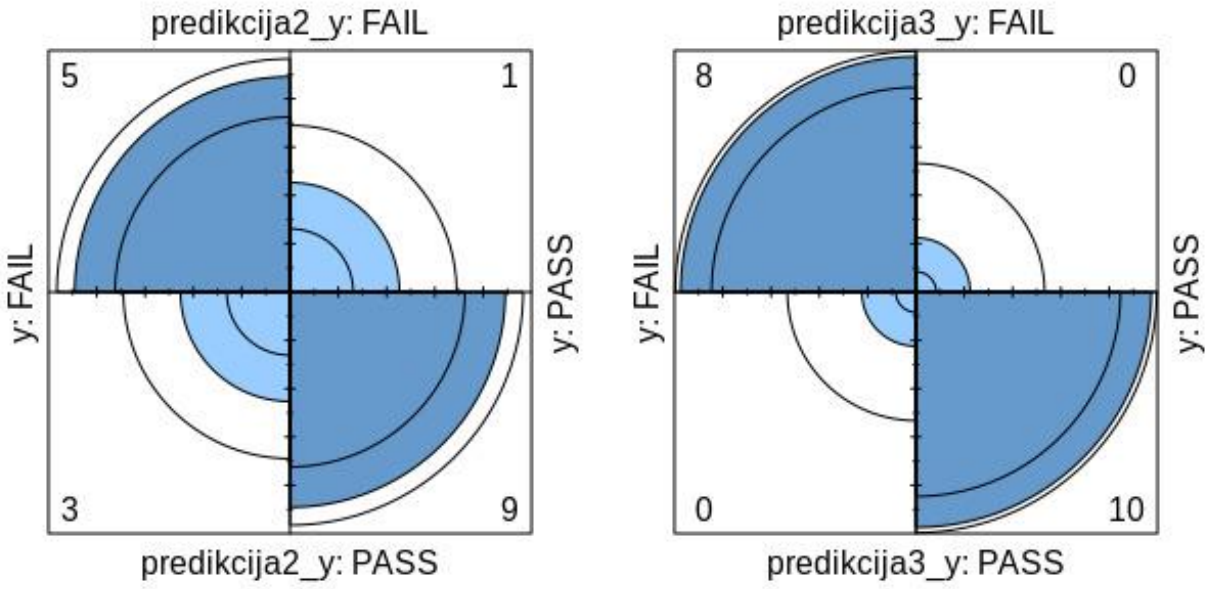
d) Najvažniji rezultati

Vezano uz drugo poglavlje ovog rada, vidimo da je najveći utjecaj varijable *„stog”* na konačnu ocjenu studenata bio u kombinaciji sa varijablama *„videos” i „lectures”,* gdje vidimo da su skoro svi studenti (uz jednu iznimku) koji su pogledali prezentaciju Stog prošli kolegij. U kombinaciji sa varijablama *„labs”, „quizzes” i „selfassesm”,* varijabla *„stog”* nije imala velik utjecaj na konačnu ocjenu studenta koliko je imala druga pridružena varijabla. Zaključak: kada gledamo utjecaj varijabli uz prezentaciju Stog, studenti koji su gledali snimljena predavanja i koji su imali najviše bodova u okviru aktivnosti ostvarili su najbolji uspjeh.



Slika 20: Grafikoni sa najboljim utjecajem varijable stog

Vezano uz treće poglavlje ovog rada, najbolji logistički model sa postotkom greške 0% je treći model u kojem je korištena kombinacija prediktora *„selfassesm”, „stog”, „lectures”, „labs” i „quizzes”.* Važno je također napomenuti kako je model sa samo dva prediktora („*stog” i „labs”*) pogrešno predvidio samo jednu vrijednost (četvrti model). Najveći postotak pogreške imao je drugi model eksperimenta u kojem su korišteni prediktori „*stog” i „selfassesm”.*



Slika 21: Usporedba najgoreg i najboljeg ligostičkog modela

Izvori

1) https://moodle.srce.hr/2019-2020/mod/resource/view.php?id=990848 (druga zadaća uz predavanja)
2) https://www.rdocumentation.org/packages/SECFISH/versions/0.1.7/topics/GLM [07.04.2020]
3) https://www.rdocumentation.org/packages/ggplot2/versions/3.3.0/topics/ggplot [07.04.2020]
4) https://www.theanalysisfactor.com/r-glm-plotting/ [07.04.2020]
5) https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html [07.04.2020]
6) https://plotly.com/r/line-and-scatter/ [07.04.2020]
7) https://stat.ethz.ch/R-manual/R-patched/library/graphics/html/fourfoldplot.html [07.04.2020]