Prozni izveštaj za projekat

Ovaj projekat je rađen sa ciljem otkrivanja faktora uspešnosti imunizacije protiv besnila. Podaci su prikupljeni u Pasterovom zavodu – Nacionalnoj referentnoj laboratoriji za besnilo. Podaci se nalaze u fajlu .xlsx, a anonimnost pacijenata je omogućena jer baza ne sadrži nikakve lične informacije. Podaci su prikupljani u periodu od 2017. do 2019. godine. Baza sadrži 17 nezavisnih varijabli relevantnih za istraživanje i 2 zavisne varijable. Nezavisne varijable su: godina, pol, godina rođenja, ustanova pošiljalac uzorka, trajanje vakcinacije, razlika završetka vakcinacije i početka ispitivanja, razlika početka vakcinacije i početka ispitivanja, razlika završetka vakcinacije i bustera, razlika prijema uzorka i uzorkovanja, razlika prijema uzorka i početka ispitivanja, razlika uzorkovanja i početka ispitivanja, ukupno dana od vađenja krvi, davanje seruma, broj jedinica, lokacija ozleda, broj ozleda i životinja koja je načinila ozledu. Zavisne varijable su: uspešnost imunizacije i rezultat analize.

Podaci su uz pomoć read\_excel() funkcije uneti u R i varijable su izabranje. Uz pomoc stringr i dplyr paketa podaci su sređivani kako bi se uklonili pogrešni unosi. Istaknute funkcije u ovom delu projekta jesu [grepl()](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/grep) i [replace()](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/replace).

Nakon uklanjanja neodgovarajućih unosa u varijable od interesa kreirana je funkcija pod nazivom remove\_outliers:

remove\_outliers <- function(x, na.rm = TRUE, ...) {

qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)

H <- 1.5 \* IQR(x, na.rm = na.rm)

y <- x

y[x < (qnt[1] - H)] <- NA

y[x > (qnt[2] + H)] <- NA

y

}

Funkcija je kreirana za svrhu uklanjanja outliera, jedan argument koji je potrebno uneti i jedan default argument. Argument koji se unosi jeste pojedinačna varijabla kojoj je potrebno ukloniti outliere a oni se definišu kao vrednosti koje se nalaze preko jedne i po vrednosti interkvartilnog raspone te varijable, sa obe strane distribucije. Argument na.rm = TRUE omogućava da se vrednosti računaju iako u varijabli ima nedostajućih podataka (na).

Nakon uklanjanja outliera sledi eksplorativna analiza podataka gde je korištena biblioteka psych, sa osvrtom na [describe()](https://www.rdocumentation.org/packages/psych/versions/2.2.3/topics/describe) funkciju koja je vrlo korisna za dobijanje deskriptivne statistike svake pojedinačne varijable ubacivanjem cele baze kao argumenta u funkciju.

Biblioteka ggplot2 se koristila za vizualizaciju podataka sa najistaknutijim funkcijama geom\_bar() i geom\_histogram() koje su služile za vizualizaciju pojedinačnih varijabli.

Shapiro-Wilk test([shapiro.test()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/shapiro.test)) se koristio da testira normalnost distribucija varijabli koje su kontinuirane.

Za inferencijalnu statistiku između zavisne i nezavsne varijable korišteni su χ2 test, t test(ili neparametrijska alternativa Wilcoxon, ukoliko uslovi za t test nisu zadovoljeni), ANOVA(ili neparametrijska alternativa Kruskal – Wallis ukoliko uslovi nisu zadovoljeni), Pirsonov produkt moment koeficijent korelacije. Funkcije primenjene za ove testove mogu se videti u Tabeli 1.

|  |  |
| --- | --- |
| Test | Funkcija |
| χ2 test | [chisq.test()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/chisq.test) |
| t test | [t.test()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test) |
| Wilcoxon | [wilcox.test()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test) |
| ANOVA | [aov()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/aov) |
| Kruskal – Wallis | [kruskal.test()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kruskal.test) |
| Pirsonov produkt moment koeficijent korelacije | [cor.test()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/cor.test) |

Tabela 1. Prikaz odgovarajućih funkcija za određeni test.

Značajni rezultati se pronalaze na:

* Wilcoxon testu između varijable godina rođenja i uspešnosti imunizacije:

W = 32710, p < .01

* χ2 testu između ustanove pošiljalaca uzorka i uspešnosti imunizacije:

χ2(26) = 40.504, p = 0.03479

* Wilcoxon testu između broja jedinica i uspešnosti imunizacije:

W = 10996, p = 0.02538

* Pirsonovoj produkt moment korelaciji između rezultata analize i godine rođenja:

r = 0.176, p < .01

* Kruskal-Wallis testu rezultata analize i ustanove pošiljalaca uzorka:

χ2(26) = 55.251, p < .01

* Wilcoxon testu između varijable rezultat analize i dat serum:

W = 7630, p < .01

accuracy precision recall f1

0.8292683 0.8750000 0.9333333 0.9032258

accuracy precision recall f1

0.7642276 0.8392857 0.8952381 0.8663594

accuracy precision recall f1

0.8536585 0.8717949 0.9714286 0.9189189

Za određivanje faktora rizika koristila se logistička regresija sa uspešnošću imunizacije kao zavisnom varijablom, a nezavisne varijable korištene su kao prediktori. Cilj je bio određivanje najboljeg modela a iterativnim procesom prediktori su ubacivani i izbacivani kako bi se odredio model sa najboljim metrikama.

Za potrebe sprovođenja logističke regresije korištena je funkcija [glm()](https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm), sa argumentom family = binomial(link = “logit”) kako bi se na generalizovanom linearnom modelu sprovodila logistička regresija. Biblioteka cater koristila se zbog funkcije [createDataPartition()](https://www.rdocumentation.org/packages/caret/versions/6.0-91/topics/createDataPartition) kako bi se postojeća baza podataka razdvojila na set za treniranje i set za testiranje modela.

Varijablu uspešnosti imunizacije bilo je potrebno iz karakter varijable promeniti u numeričku I to se vršilo pomoću funkcije [replace()](https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/replace).

Funkcija compute.eval.metrics() kreirana je kako bi se računala metrika modela.

compute.eval.metrics <- function(cm) {

TP <- cm[2,2]

TN <- cm[1,1]

FP <- cm[1,2]

FN <- cm[2,1]

acc <- (TP + TN) / (TP + TN + FP + FN)

prec <- TP / (TP + FP)

recall <- TP / (TP + FN)

f1 <- 2 \* (prec \* recall) / (prec + recall)

c(accuracy = acc, precision = prec, recall = recall, f1 = f1)

}

Ova funckija prima jedan argument koji mora biti matrica zabune(confusion matrix). Matrica zabune se dobija tako što se pravi tabela kontigencije od pravih vrednosti iz seta za testiranje i vrednosti koje model predviđa. Output ove funkcije jesu četiri metrike: tačnost(accuracy), preciznost(precision), odziv(recall) i f mera(f1).

Konačni model koji se usvojio kao najbolji bio je model sa jednim prediktorom(godina rođenja). Metrike ovog modela su sledeće: tačnost je iznosila 0.869, preciznost 0.867, odziv 1, a f mera 0.929.

Faktor rizika koji najbolje predviđa neuspešnost imunizacije jesu godine rođenja i kako godine rođenja rastu, rizik za neuspešnu imunizaciju opada. Radi lakše interpretacije, varijablu godina rođenja trebalo je oduzeti od trenutne godine kako bi se dobila starost pacijenata a tada bi interpretacija modela bila sledeća: sa porastom starosti raste i verovatnoća da će pacijent biti neuspešno imunizovan.