Analiza podataka tumori

Podaci o tumorima preuzeti su sa stranice UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php) u skupu pod nazivom "Breast Cancer Wisconsin (Original)". Podaci su podijeljeni u dvije datoteke: `.data` i `.names`. Prva sadrži podatke, dok druga sadrži nazive varijabli.

1. Učitajte podatke iz datoteke `.data` u data frame naziva **bc**.

```
bc <- read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data", sep = ",", na.strings = c("?"))
```

2. Koliko instanci i varijabli sadrži skup podataka? Koje su vrste varijabli s obzirom na mjernu skalu?
3. Dodajte nazive stupaca prema datoteci `.names` u data frame **bc**. Izbacite prvi stupac (ID).
4. Koliko ima nedostajućih vrijednosti u skupu podataka? Izračunajte udio nedostajućih vrijednosti za svaku varijablu.
5. Zamijenite nedostajuće vrijednosti s medijanom.

Podatke ćemo koristiti za klasifikaciju tkiva na temelju određenih karakteristika u dvije klase: benigni tumor ili maligni tumor. Zavisna (izlazna) varijabla bit će **class**.

6. Pretvorite varijablu **class** u kategoričnu. Označite vrijednosti oznakama „B“ za benigni i „M“ za maligni tumor.
7. Podijelite podatke slučajnim odabirom u omjeru 70:30 na dva skupa: **trening** i **test**.
8. Izračunajte udio pojedine klase u sva tri skupa: **bc**, **trening** i **test**.

Podatke iz skupa **trening** koristit ćemo za izgradnju modela. Podatke iz skupa **test** koristit ćemo za provjeru modela.

9. Napravite model logističke regresije s **class** kao zavisnom varijablom, a sve preostale varijable kao nezavisne.
10. Protumačite koeficijent uz varijablu **clump**.
11. Koristite model za predviđanje na podacima iz skupa **test**. Izvršite klasifikaciju podataka koristeći prag 0.5, tako da sve vrijednosti iznad praga budu klasificirane kao maligni tumor.
12. Napravite konfuzijsku matricu. Koliko iznosi točnost klasifikacije?