Ime i prezime: _____ Matični broj: _____

Ispitna pitanja

Dan je skup podataka koji sadrži različite informacije o pacijentima koji imaju opasnost od zatajenja srca. Skup podataka sadrži 8 redaka i 8 stupaca. Riješite sljedeće zadatke koristeći zadani skup podataka. **Nije potrebno** zapisivati rezultate svih međukoraka pri izračunu, dovoljno je prikazati upotrebu formule i konačnu vrijednost.

	Dob	Spol	Pušac	Dijabetes	Visoki krvni tlak	Kreatinin (mg/dL)	Natrij (mEq/L)	Zatajenje srca
0	20	M	DA	0	1	0.5	120	0
1	50	F	NE	1	0	1.1	130	1
2	30	F	NE	0	0	1.2	150	0
3	40	M	DA	0	1	0.9	140	0
4	60	M	DA	1	0	1.3	120	1
5	30	F	NE	0	1	1.8	150	1
6	20	F	NE	1	0	2.5	140	0
7	50	F	DA	0	1	3.5	160	1

1. **[8 bod.]**

   1.1. **[4 bod.]** Pomoću dijagrama pravokutnika (*box and whisker plots*) usporedite razinu natrija kod muškaraca i žena. Koji je raspon razine natrija, a kolika je medijalna razina natrija po spolu?

   1.2. **[4 bod.]** Pomoću prikladnog dijagrama prikažite odnos između dobi pacijenta i razine natrija. Koliko iznosi koeficijent linearne (Pearsonove) korelacije između ove dvije varijable? Formula: $r = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_i^n(X_i-\bar{X})^2 \cdot \sum_i^n(Y_i-\bar{Y})^2}}$.
   Napomena: smijete zaokružiti prosječne vrijednosti za obje varijable na cijeli broj.

2. **[11 bod.]**

   2.1. **[3 bod.]** Pomoću prikladnog dijagrama prikažite distribuciju dobi pacijenata. Koliko je pacijenata mlađe, a koliko starije od prosječne dobi pacijenta?

   2.2. **[4 bod.]** Pomoću Kolmogorov-Smirnov testa testirajte je li dob distribuirana prema normalnoj distribuciji, gdje je μ jedan prosječnoj dobi, a σ standardnoj devijaciji dobi. Funkcija kumulativne teorijske distribucije jest:

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(\frac{(x-\mu)^2}{2\sigma^2})}.$$

   Kritična vrijednost KS testa za uzorak veličine 8 i $p = 0.05$ je 0.457.

   2.3. **[4 bod.]** Testirajte sljedeću nultu hipotezu upotrebom **jednostranog** binomijalnog testa: *Vjerojatnost zatajenja srca kod pušača iznosi* 10%. Alternativna hipoteza pretpostavlja veću vjerojatnost (> 10%) zatajenja srca kod pušača, a testna statistika se može odrediti iz skupa podataka. Funkcija binomijalne distribucije jest:

$$P_x = \binom{n}{x} p^x q^{n-x},$$

   gdje je x broj pojavljivanja određenog događaja nakon n ponavljanja.

Moguće je skupiti ukupno 50 bodova.
Za prolazak je potrebno skupiti 25 bodova.
Ispit se piše 120 minuta.

3. **[10 bod.]**

3.1. **[3 bod.]** Pomoću modela linearne regresije modelirajte dob pacijenta **samo za muškarce** pomoću sljedećih prediktora: dijabetes i visoki krvni tlak. Procijenite parametre modela pomoću metode najmanjih kvadrata, odnosno izračunajte $\hat{\beta} = (X^T X)^{-1} X^T Y$.

3.2. **[3 bod.]** Primijenite logaritamsku transformaciju nad izlaznom varijablom (*Dob*). Ponovno procijenite parametre modela pomoću metode najmanjih kvadrata, ovaj puta s logaritmiranom izlaznom varijablom.

3.3. **[2 bod.]** Za varijable visoki krvni tlak i dijabetes odredite koliki je njihov aditivni učinak u prvom modelu, odnosno multiplikativni učinak u drugom modelu. Kako se mijenja vrijednost izlazne varijable u prvom, a kako u drugom modelu, ako pacijent ima visoki krvni tlak ili dijabetes?

3.4. **[2 bod.]** Koji od dva modela bolje modelira dob **muških** pacijenta? Odgovorite na pitanje upotrebom koeficijenta $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$.

4. **[11 bod.]**

4.1. **[4 bod.]** Pomoću algoritma 1-NN klasificirajte pacijente prema riziku od zatajenja srca na temelju značajki: dijabetes i visoki krvni tlak. Prvo podijelite skup podataka na podskupove za trening i test tako da prvih 5 redaka u skupu podataka čini podskup za trening, a iduća 3 retka podskup za test. Nacrtajte matricu zabune za algoritam 1-NN na podskupu za test.

4.2. **[5 bod.]** Pomoću algoritma stablo odluke (DT) također klasificirajte pacijente prema riziku od zatajenja srca na temelju značajki: dijabetes i visoki krvni tlak. Neka maksimalna dubina stabla bude jednaka 2, a neka se kao mjera nečistoće čvora koristi Gini indeks. Vrijednost Gini indeksa jest:

$$Gini = 1 - \sum_{i=1}^{m} P(i)^2,$$

gdje je m broj klasa, a $P(i)$ vjerojatnost pojave primjerka s oznakom klase i u čvoru. Napomena: manja vrijednost za mjeru nečistoće je bolja. Nacrtajte matricu zabune za algoritam DT na podskupu za test.

4.3. **[2 bod.]** Koji od algoritama 1-NN i DT ima bolje performanse na podskupu za test?

5. **[10 bod.]**

5.1. **[7 bod.]** Zanima nas mogu li se pacijenti općenito grupirati u grupe na temelju njihove razine kreatinina u krvi. Pomoću metode lakta pronađite najprikladniju vrijednost za parametar k u algoritmu k-means te zatim grupirajte sve primjerke u skupu podataka na temelju ove značajke. Napomena: za "nasumično" odabrani centar neke grupe u algoritmu k-means uvijek uzmite prvi idući redak u skupu podataka (odnosno, prvi redak za prvi centar, drugi redak za drugi centar, itd.). Algoritam k-means je potrebno vrtiti samo 2 iteracije.

5.2. **[3 bod.]** Vizualizirajte grupe pomoću prikladnog dijagrama u prostoru određenom ovom značajkom.

Moguće je skupiti ukupno 50 bodova.
Za prolazak je potrebno skupiti 25 bodova.
Ispit se piše 120 minuta.