Istraživanje podataka - praktični deo ispita, jun 2021. g.
smer Informatika

Broj indeksa	**Ime i prezime**

Broj poena po zadacima je:

Zadatak	1	2	3	4	5	**Zbir**
maks	8	15	12	30	35	**100**
Osvojeno						

1. Za sledeće atribute, navesti koje su vrste: dodeljeni broj trkaču na maratonu prilikom prijave i vreme za koje je učesnik istrčao maraton. Obrazložiti odgovor.

2. U tabeli su date posteriorne verovatnoće dobijene primenom modela klasifikacije na test podatke. Skup podataka ima dve klase. Nacrtati ROC krivu na osnovu zadatih podatka i izračunati AUC. Šta se može zaključiti o ponašanju algoritma na osnovu ROC krive?

Instanca	Klasa	$P(+ \mid X)$
1	-	0,1
2	-	0,7
3	+	0,4
4	-	0,2
5	+	0,8
6	+	0,9

3. Nad datim skupom transakcija primeniti Apriori algoritam za računanje čestih skupova stavki. Nacrtati mrežu čestih skupova koje Apriori može razmatrati i jednom crtom precrtati one koji se odsecaju nakon računanja podrške, a dva puta one za čije odsecanje nije potrebno računanje podrške. Za česte skupove izračunati podršku i ispisati je u gornjem desnom uglu odgovarajućeg čvora. Za zahtevanu podršku uzeti vrednost 0,3.

1	$\{A\}$
2	$\{C\}$
3	$\{C, D\}$
4	$\{A, B, C\}$
5	$\{A, C, D\}$
6	$\{A, C\}$
7	$\{A, B\}$
8	$\{B\}$
9	$\{A, B, D\}$
10	$\{A, B, C, D\}$

4. Na Desktopu u direktorijumu **ipJun22021_skupovi** nalazi se skup podataka *klasifikacija_vina.csv* sa podacima o vinima. Primenom alata IBM SPSS Modeler izvršiti klasifikaciju nad skupom. Ciljni atribut je kolona *type*. U radnom toku uraditi i odgovoriti na pitanja:

- Eliminisati slogove koji sadrže elemente van granica određene metodom sa kvartilima.
- Primeniti algoritam C5.0 i zadati da je minimalan broj instanci koji mora da bude u dete-čvoru 10. Dobijeni model nazvati *model1*.
- Primeniti algoritam SVM sa linearnim kernelom. Dobijeni model nazvati *model2*.
- Koji atributi su najznačajniji za pravljenje *model1*.
- Diskutovati i uporediti *model1* i *model2*.

Podatke o dobijenim modelima (preciznost i matrice konfuzije na trening i test skupu) sačuvati u html datotekama.

Radni tok eksportovati i dodeliti mu ime u formatu **klasifikacija_*vasBrojIndeksa***. Odgovore pišite u datoteku sa nazivom **klasifikacija_*vasBrojIndeksa_odgovori*.txt**.

5. Na Desktopu u direktorijumu **ipJun22021_skupovi** nalazi se skup podataka *klasterovanje.csv* sa 4 numerička atributa. Koristeći skup i biblioteke programskog jezika Python izvršiti hijerarhijsko klasterovanje.

   U programu:

   - Primeniti klasterovanje korišćenjem svih atributa i napraviti modele za 5, 6 i 7 klastera za različite načine spajanja klastera.
   - Svaki dobijeni model primeniti nad instancama i rezultat klasterovanja prikazati pomoću grafika sa razbacanim elementima (eng. scatter). Pre prikazivanja rezultata grafički, primenom tehnike PCA smanjiti broj atributa na dimenziju 2 i dobijene atribute koristiti za grafički prikaz. Svakom klasteru dodeliti jedinstvenu boju i označiti koja je veza korišćena i koliki je senka koeficijent za klasterovanje.

   U komentarima odgovoriti na pitanja:

   - Da li je bilo obrade podataka pre klasterovanja? Zašto?
   - Ukratko napisati zaključke o dobijenim modelima. Koji tip veze kod hijerarhijskog klasterovanja daje najbolje rezultate?
   - Koliko početne varijanse u skupu je objašnjeno sa dva atributa nakon primene PCA?

   Skriptu/datoteci dodeliti ime u formatu **klasterovanje_*vasBrojIndeksa***. Ako koristite Jupyter Notebook, dokument sačuvajte sa ekstenzijom *ipynb* i kao html dokument. Odgovore pišite u datoteku sa nazivom **klasterovanje_*vasBrojIndeksa_odgovori***. Ukoliko pišete skript u pj Python, izlaz programa sačuvajte u datoteci sa nazivom u formatu **izlaz_*vasBrojIndeksa*.txt** i sačuvajte sliku.

Uputstvo za čuvanje rada: Na Desktopu napravite direktorijum sa nazivom u formatu **ip.jun2.2021.*ime.prezime.brojIndeksa*** gde umesto ime, prezime i broj indeksa stavite Vaše podatke. Npr, `ip.jun2.2021.petar.petrovic.543_2014` U tom direktorijumu čuvajte rešenja praktičnih zadataka i datoteke sa odgovorima.