Istraživanje podataka 1 - pismeni deo ispita, jun 2020. g.

Broj indeksa	**Ime i prezime**

Zadaci se rade 150 minuta. Broj poena po zadacima je:

Zadatak	1	2	3	3	3	**Zbir**
maks	6	15	14	32	33	**100**
Osvojeno						

1. Za sledeće atribute, navesti koje su vrste: površina tela i stepen obrazovanja. 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. Šta se može zaključiti o ponašanju algoritma na osnovu ROC krive?

Instanca	Klasa	$P(+ \mid X)$
1	+	0,2
2	-	0,05
3	+	0,35
4	-	0,25
5	+	0,65

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 računanje podrške nije potrebno. Za zahtevanu podršku uzeti vrednost 30%. Izračunati podršku čestih skupova stavki.
   - Izračunati pouzdanost za sledeća pravila i odrediti koje pravilo je bolje prema pouzdanosti: $\{A, D\} \rightarrow \{C\}$ ili $\{A, C\} \rightarrow \{D\}$.

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

4. Na Desktopu u direktorijumu **ipJun2020** nalazi se skup podataka *klasifikacija.csv* sa podacima o vinima. Primenom alata IBM SPSS Modeler izvršiti klasifikaciju nad skupom. Ciljni atribut je kolona *class*. U radnom toku uraditi i odgovoriti na pitanja:

- Primeniti algoritam C5.0. Dobijeni model nazvati *model1*.
- Koji atributi su najznačajniji za pravljenje *modela1*?
- Komentarisati model *model1*. Koja je dubina drveta odlučivanja?
- Primeniti algoritam C5.0 sa definisanom matricom cena tako da se dobro klasifikuju instance klase 4. Dobijeni model nazvati *model2*.
- Komentarisati model *model2*.
- Primeniti analizu glavnih komponenti nad skupom radi smanjenja broja atributa.
- Na koliko atributa ste smanjili skup? Obrazložiti odgovor.
- Primeniti algoritam C5.0 nad transformisanim skupom i dobijeni model nazovite *model3*.
- Komentarisati model *model3*.

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 pisati u datoteku sa nazivom **klasifikacija_*vasBrojIndeksa_odgovori*.txt** ili u okviru radnog toka kao komentare.

5. Na Desktopu u direktorijumu **ipJun2020** nalazi se skup podataka *klasterovanje.csv* sa 2 numerička atributa *a* i *b*. Koristeći skup i biblioteke programskog jezika Python izvršiti hijerarhijsko klasterovanje za različit broj klastera u intervalu od 3 do 13 i primenom različitih veza za određivanje bliskosti dva klastera. Kao meru rastojanja koristiti Euklidsko rastojanje.

   Za svaku primenjenu vezu:

   - rezultat kalsterovanja prikazati pomoću grafika sa silueta koeficijentom i brojem izdvojenih klastera
   - odrediti najbolji broj klastera prema silueta koeficijentu
   - za najbolji broj klastera prema silueta koeficijentu, rezultat klasterovanja instanci prikazati pomoću grafika sa razbacanim elementima (eng. scatter). Svakom klasteru dodeliti jedinstvenu boju i označiti koja je veza korišćena i koliki je senka koeficijent za klasterovanje.

   Sačuvati dobijene slike u png formatu.

   U komentarima odgovoriti na pitanja:

   - Da li je bilo obrade podataka pre klasterovanja? Zašto?
   - Ukratko napisati zaključke o izvršenim klasterovanjima.

   Skript sačuvati i dodeliti mu ime u formatu **klasterovanje_*vasBrojIndeksa***. Odgovore pisati u datoteku sa nazivom **klasterovanje_*vasBrojIndeksa_odgovori*.txt**.

Uputstvo za čuvanje rada: Na Desktopu napravite direktorijum sa nazivom u formatu **ip.Jun.2020.*ime.prezime.brojIndeksa*** gde umesto ime, prezime i broj indeksa stavite Vaše podatke. Npr, `ip.Jun.2020.petar.petrovic.543_2014` U tom direktorijumu čuvajte rešenja praktičnih zadataka i datoteke sa odgovorima.