Istraživanje podataka - praktični deo ispita, jul 2019.

Broj indeksa	**Ime i prezime**

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

Zadatak	1	2	3	**Zbir**
maks	30	30	40	**100**
Osvojeno				

1. Na Desktopu u direktorijumu **ipJul2019** nalazi se skup podataka *atelje212.csv* sa podacima o glumcima koji igraju u predstavama pozorišta Atelje 212. Primenom alata IBM SPSS Modeler i algoritma Apriori pronaći pravila pridruživanja o glumcima.

   - Pronaći pravila pridruživanja o glumcima koji zajedno igraju u predstavama. Postaviti uslove da je najmanja podrška za telo 20%, a pouzdanost pravila 60%. Dobijeni model nazvati *model1*.
   - Koja pravila pridruživanja u modelu 1 su najzanimljivija? Zašto?
   - Na osnovu modela 1 odgovoriti na pitanje: Ako u predstavi glumi Jelena Đokić, za koga se još može očekivati da će glumiti u predstavi? Pravila na osnovu kojih se donosi zaključak sačuvati kao poseban model i nazvati ga *modelJDj*, a sačuvati ih i u html datoteci *JDj*.
   - Pronaći pravila pridruživanja o glumcima uzimajući u obzir i kada glumac ne igra u predstavi. Postaviti uslove da je najmanja podrška za telo 20%, a pouzdanost pravila 60%. Dobijeni model nazvati *model2*.
   - Na osnovu modela 2 odgovoriti na pitanje: Ako u predstavi glumi Katarina Žutić, za koga se može očekivati da neće glumiti u predstavi? Pravila na osnovu kojih se donosi zaključak sačuvati u html datoteci *KZ*.

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

2. Na Desktopu u direktorijumu **ipJul2019** nalazi se skup podataka *zitarice_klasifikacija.csv* sa podacima o žitaricama. Primenom alata IBM SPSS Modeler izvršiti klasifikaciju nad skupom. Ciljni atribut je kolona *class*. U radnom toku uraditi i odgovoriti na pitanja:

   - Primeniti algoritam drveta odlučivanja C&RT. Samostalno izabrati parametre. Dobijeni model nazvati *model1*.
   - Primeniti algoritam K najbližih suseda. Postaviti da se izabere najbolji broj suseda iz intervala $[3, 8]$ i da se izabere 5 najboljih atributa. Izvršiti normalizaciju atributa i primeniti Euklidsko rastojanje. Dobijeni model nazvati *model2*.
   - Koje parametri su postavljeni pri pravljenju *modela1*? Zašto?
   - Koji atributi su najznačajniji za pravljenje *modela1*?
   - Koji atributi su izabrani pri pravljenju *modela2*?
   - 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 pisati u datoteku sa nazivom **klasifikacija_*vasBrojIndeksa_odgovori*.txt**.

3. Na Desktopu u direktorijumu **ipJul2019** nalazi se skup podataka *zitarice_klasterovanje.csv*. Koristeći skup i biblioteke programskog jezika Python izvršiti hijerarhijsko klasterovanje.

   U programu:

   - Izvršiti klasterovanje korišćenjem svih numeričkih atributa i najmanje, najduže i prosečne veze. Rezultate klasterovanja za sve veze prikazati pomoću dendograma. Sve dendograme prikazati na jednoj slici i za svaki kao naslov ispisati ime korišćene veze pri klasterovanju. Sliku sačuvati u png formatu i nazvati *dendogrami*.
   - Na osnovu dendograma odrediti koja je veza za spajanje klastera najpogodnija za klasterovanje datog skupa i koji broj klastera bi trebalo izabrati. Za izabranu vezu spajanja i broj klastera svakoj instanci odrediti klaster kom pripada. Ispisati senka kofecijent za klasterovanje sa izabranim brojem klastera.
   - Rezultat klasterovanja instanci sa izabranim brojem klastera 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. Dobijenu sliku sačuvati u png formatu.

   U komentarima odgovoriti na pitanja:

   - Da li je bilo obrade podataka pre klasterovanja? Zašto?
   - Koji tip veze kod hijerarhijskog klasterovanja je najpogodniji za dati skup? Obrazložiti odgovor.
   - Koji broj klastera ste izabrali? Zašto?

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

Uputstvo za čuvanje rada: Na Desktopu napravite direktorijum sa nazivom u formatu **ip.Jul.2019.*ime.prezime.brojIndeksa*** gde umesto ime, prezime i broj indeksa stavite vaše podatke. Npr, `ip.Jul.2019.petar.petrovic.543_2014` U tom direktorijumu čuvajte rešenja zadataka i datoteke sa odgovorima.

Opis atributa skupa *zitarice*:

- *name*: ime žitarice
- *class*: klasa (samo za klasifikaciju)
- *calories*: kalorija
- *protein*: grama proteina
- *fat*: grama masti
- *sodium*: miligram natrijuma
- *fiber*: grama dijetetskih vlakana
- *carbo*: grama složenih ugljenih hidrata
- *sugars*: grama šećera
- *potass*: miligrami kalijuma
- *vitamins*: vitamini i minerali