Istraživanje podataka - primer pismenog dela ispita

1. Skup podataka *oscar.csv* sadrži podatke o tvitovima koji se odnose na dodelu Oskara 2017. Koristeći alat IBM SPSS Modeler izdvojiti pravila pridruživanja o rečima iz tvitova. Prilikom pravljenja modela zadati sledeće uslove:

   - najmanja podrška tela je 3%.
   - najmanja pouzdanost je 70%.

   U okviru toka uraditi:

   - Na osnovu dobijenog modela napraviti novi model *najboljaPravilaLift* koji sadrži pravila koja su nabolja prema Lift meri.
   - U html datoteku izdvojiti pravila koja ne sadrže reč *oscar*.

   Odgovoriti na pitanja:

   - Koliko je pravila pridruživanja u modelu?
   - Kolika je najmanja, a kolika najveća podrška pravila?
   - Koje pravilo je najzanimljivije po Lift meri? Objasniti zašto.
   - Pronaći najbolje pravilo prema Lift meri za drugu transakciju u skupu kada se stavka koja je u glavi
     - javlja u transakciji.
     - ne javlja u transakciji.

   Radni tok eksportovati i rezultat imenovti u formatu **SPSS_pravila_*vasBrojIndeksa***. Odgovore pišite u datoteku sa nazivom **SPSS_pravila_*vasBrojIndeksa*_*odgovori***.

2. Koristeći skup podataka *zoo.csv* izvršiti klasifikaciju algoritmom K najbližih suseda primenom unakrsne validacije u programskom jeziku Python. Uraditi redom:

   - Napraviti različite modele klasifikacije promenom: broja suseda, mere rastojanja i težine suseda. Izdvojiti izveštaj o klasifikaciji za najbolji model prema preciznosti i prikazati matricu konfuzije za trening i test skup.
   - Primeniti tehniku PCA radi smanjenja dimenzija skupa podataka.
   - Nad skupom dobijenim nakon primene PCA napraviti različite modele klasifikacije promenom: broja suseda, mere rastojanja i težine suseda. Izdvojiti izveštaj o klasifikaciji za najbolji model prema preciznosti i prikazati matricu konfuzije za trening i test skup. Prikazati grafički test podatke pomoću grafika sa razbacanim elementima (eng. scatter). Koristiti 2 atributa koja nose najviše informacija o promenljivosti u skupu.

   U komentarima:

   - Koji atributi iz zadatog skupa nisu korišćeni i zašto?
   - Na koji broj atributa je smanjen skup korišćenjem PCA tehnike i zašto?
   - Opisati dobijene modele i uporediti ih.

Skriptu dodeliti ime u formatu **klasifikacija_*vasBrojIndeksa***. Izlaz programa sačuvajte u datoteci sa nazivom u formatu **izlaz_*vasBrojIndeksa*.txt**. Odgovore pišite u datoteku sa nazivom **klasifikacija_*vasBrojIndeksa*_odgovori**.

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

3. Za sledeće atribute navesti koje su vrste: ime sporta i vreme za koje je učesnik istrčao maraton. Obrazložiti odgovor.

4. Dat je skup instanci I1-I6. Izvršiti nad njima hijerarhijsko klasterovanje korišćenjem Menhetn rastojanja i *max* veze. Rezulatat prikazati dendogramom. Ako je prag za spajanje klastera 8, identifikovati klastere koji bi bili izdvojeni.

Instanca	X	Y	Z
I1	3	4	2
I2	5	5	-2
I3	-3	-1	-2
I4	-2	4	-3
I5	-2	3	-4
I6	4	2	3

5. Data je izveštaj za izvršenu klasifikaciju na trening i test skupu:

Results for output field Class

Comparing $C-Class with Class

'Partition'	Testing		Training	
Correct	173	78.64%	425	80.49%
Wrong	47	21.36%	103	19.51%
Total	220		528	

Coincidence Matrix for $C-Class (rows show actuals)

'Partition' = Testing	0	1
0	159	10
1	37	14

'Partition' = Training	0	1
0	384	17
1	86	41

- Da li je došlo do preprilagođavanja podacima za treniranje? Obrazložiti odgovor.
- Cilj je napraviti model koji dobro klasifikuje instance klase **1**, čak i po cenu da instance klase **0** budu lošije klasifikovane. Definisati matricu cene za taj cilj.

Dodatni zadatak za vežbu

Skup podataka *diamonds.csv* sadrži podatke o dijamantima. Koristeći skup i alat IBM SPSS Modeler izvršiti klasterovanje nad skupom primenom algoritma K-sredina za 4 klastera.

U komentarima odgovoriti na pitanja:

- Koliki je kvalitet dobijenih modela?
- Koji atributi su najznačajniji za pravljenje modela?
- Uporediti najveće klastere iz dobijenih modela.