IP teorija - sept2 2020.

1. Šta sve spada u pretprocesiranje podataka, i zašto se podaci uopšte pretprocesiraju?
2. Dat je box-plot dijagram na kome su prikazani atributi iz skupa IRIS - sepalwidth, sepallength, petalwidth, petallength. Treba diskutovati o stvarima koje se mogu zaključiti sa dijagrama (da li imaju smisla, gde ima elemenata van granica, da li su vrednosti ravnomerno raspoređene...)
3. Na koje sve načine se može uzorkovati?
4. Šta je potkresivanje i kada se koristi? Navesti sve algoritme klasifikacije drvetom odlučivanja koje znate i reći koji od njih koriste potkresivanje.
5. Detaljno opisati algoritam razdvajajućeg hijerarhijskog klasterovanja.
6. Detaljno opisati algoritam klasterovanja k-sredina. Na koje sve načine je moguće birati centroide na početku ovog procesa?
7. Detaljno opisati vertikalni apriori algoritam.
8. Šta je tabela kontigenata? Dati primer tabele kontigenata u kojoj podrška i pouzdanost ne daju dobre informacije, ali lift mera daje.
9. Koji problemi mogu nastati pri diskretizaciji podataka za proces nalaženja pravila pridruživanja?
10. Data je tabela frekvencija pojavljivanja reči P1..P5 u dokumentima D1...D5.
    a. Potrebno je normalizovati vektore pojavljivanja reči u dokumentu L1 normom.
    b. Odrediti podršku (P1,P2,P3) koristeći min-apriori pristup.