Istraživanje podataka 1 - vežbe 10, 2020.

1 Zadaci

1. U programskom jeziku Python izvršiti klasterovanje nad skupom *unbalance.csv* primenom algoritama K-sredina, DBSCAN i hijerarhijskog sakupljajućeg klasterovanja za različit broj klastera. Skup *unbalance.csv* sadrži numeričke atribute X i Y. Rezultate klasterovanja prikazati pomoću šeme sa raspršenim elementima. Za svaki algoritam napraviti i grafik koji prikazuje silueta koeficijent za različit broj izdvojenih klastera.

   Rešenje: *klasterovanje_ nebalansiranog_ skupa.py*

2. Izvršiti klasterovanje država na osnovu podataka o ishrani stanovništva primenom algoritma K-sredina u alatu IBM SPSS Modeler. Skup *ishrana.csv* sadrži podatak koliko su određene namirnice zastupljene u ishrani stanovništva nekih država. Vrednosti numeričkih atributa su izražene u procentima. Atributi skupa su:

   - *Country* - država
   - *RedMeat* - crveno meso
   - *Eggs* - jaja
   - *Milk* - mleko
   - *Fish* - riba
   - *Cereals* - žitarice
   - *Starch* - skrob
   - *Nuts* - koštunjavo voće
   - *Fr&Veg* - voće i povrće

   Rešenje:

   - radni tok: *drzave_ ishrana.str*
   - komentari: *ipVezbe10Primer2.pdf*

3. U programskom jeziku Python izvršiti hijerarhijsko sakupljajuće klasterovanje nad skupom u datoteci *studenti.csv* za različit broj klastera u intervalu od 2 do 34 i primenom različitih veza za određivanje bliskosti dva klastera. Skup u datoteci *studenti.csv* sadrži podatke o studentima koji su diplomirali. Za svakog studenta su izdvojeni sledeći podaci: indeks, naziv smera koji je diplomirao, dužina studiranja u godinama, broj položenih ispita i prosečna ocena. Za svaku primenjenu vezu

   - rezultat kalsterovanja prikazati pomoću grafika sa silueta koeficijentom i brojem izdvojenih klastera

- za klasterovanje sa najvećim senka koeficijentom izdvojiti deskriptivne statistike za svaki klaster da bi se uočilo šta ga karakteriše

Rešenje: *klasterovanje_studenti.py*