Zadaća br. 3

Datum objave:	**10.01.2022. god.**
Datum predaje zadaće:	**23.01.2022. god. u 23:59:59 h**
Ukupan broj bodova:	**10 bodova**

Za izradu zadaće se koristi *R Studio* okruženje i *R markdown* (Rmd) format koji se koristi na laboratorijskim vježbama. Izrada zadaće sastoji se od programskog koda (Rmd/HTML/PDF format) i dokumentacije (PDF format) sa izradom zadataka. Predaja zadaće se vrši preko Zamgera u nekom od podržanih formata za kompresovane *file*-ove (RAR, ZIP i sl.).

Zadaća se radi u timu. Timovi koji su prijavljeni za zadaću 1 i 2, ostaju isti za zadaću 3. Obratiti pažnju na zaduženja pojedinačnih članova tima. Dovoljno je da samo jedan od studenata u timu izvrši predaju zadaće putem Zamgera.

Dodatna diskusija za zadaću 3 će se održati na predavanjima 11.01.2022. godine.
U slučaju da se zadaća u potpunosti uradi prije roka zaključno sa 17.01.2022. god. i prezentira 18.01.2022. god., studenti mogu dobiti 2 nagradna boda.

Zadatak 1. (Prototip-bazirani klastering) **(Ukupno: 6 bodova)**

Radi se zajednički.

Potrebno je odabrati set podataka primjenljiv za rješavanje problema *clusteringa*. Neophodno je da odabrani set odgovara realnim podacima, te da set podataka ima sljedeće osobine:

- Broj instanci mora biti veći ili jednak 1000;
- Broj atributa mora biti veći ili jednak 15;
- Podaci moraju biti mješovitog tipa, tj. potrebno je da set podataka sadrži i numeričke i kategoričke tipove atributa.

a) Analizirati set podataka i izvršiti sljedeće:

   - Odrediti tipove i distribuciju atributa;
   - Ispuniti nedostajuće i ukloniti nepodobne vrijednosti ukoliko je potrebno;
   - Uraditi transformaciju i skaliranje atributa u skladu sa njihovom distribucijom i algoritmima koji će se koristiti u narednim koracima;
   - Odrediti da li set podataka ima klastering tendenciju;
   - Utvrditi da li je potrebno izvršiti dodatne analize za podatke, kao i dodatne transformacije, skaliranja i sl.

   (1 bod)

b) Primijeniti PAM *k-medoids* algoritam nad prethodno pripremljenim setom podataka (pritom obrazložiti sve korake pripreme koji su izvršeni, kao i njihov utjecaj na efikasnost PAM algoritma). Odabrati metriku distance i optimalnu vrijednost broja klastera *k* s kojima

se dobivaju najbolje performanse i obrazložiti njihov odabir. Evaluirati dobiveni model pomoću internih mjera validacije i objasniti dobivene rezultate. Ukoliko su dobiveni rezultati zadovoljavajući, obrazložiti zašto, a u suprotnom izvršiti određena poboljšanja. **(2 boda)**

c) Primijeniti *k-means* algoritam nad prethodno pripremljenim setom podataka (pritom obrazložiti sve korake pripreme koji su izvršeni, kao i njihov utjecaj na efikasnost *k-means* algoritma). Odabrati metriku distance i optimalnu vrijednost broja klastera *k* s kojima se dobivaju najbolje performanse i obrazložiti njihov odabir. Evaluirati dobiveni model i objasniti dobivene rezultate. Ukoliko su dobiveni rezultati zadovoljavajući, obrazložiti zašto, a u suprotnom izvršiti određena poboljšanja. **(2 boda)**

d) Smanjiti dimenzionalnost seta podataka koristeći PCA analizu, a zatim primijeniti PAM *k-medoids* i *k-means* algoritme nad rezultujućim setom podataka. Objasniti izvršeni postupak i izvršiti evaluaciju performansi dobivenog modela. **(1 bod)**

Zadatak 2. (Primjena raznih metoda klasteringa) **(Ukupno: 2 boda)**

Radi se pojedinačno.

Svaki član tima treba da odabere jedan algoritam klasteringa, pri čemu odabrani algoritmi moraju biti različiti i treba da pripadaju raznim metodama (prototip-bazirani koji nije korišten u prethodnom zadatku, hijerarhijski, zasnovan na gustoći, *fuzzy* ili baziran na modelu), što uključuje i metode koje je potrebno samostalno izučiti, tj. koje nisu obrađene u sklopu predavanja. Odabranim metodama riješiti prethodno definisani problem. Koristiti skup podataka za koji se smatra da je najpogodniji (originalni bez urađenog pod 1a), originalni sa urađenim pod 1a), sa PCA). Svaki član tima treba da opiše primijenjeni algoritam, implementira ga u R-u i izvrši njegovu evaluaciju na osnovu odgovarajućih metoda validacije.

U izvještaju je potrebno naznačiti koji algoritam je radio koji član tima, a individualno obrađeni dijelovi zadaće i izvještaja treba da dokumentuju način implementacije odabranog algoritma i njegovu primjenu, zajedno sa objašnjenjem da li član tima smatra da algoritam koji je primijenjen daje dobre rezultate.

Zadatak 3. (IEEE rad) **(Ukupno: 2 boda)**

Radi se zajednički.

Potrebno je grupno objediniti urađeno i napisati izvještaj u obliku IEEE rada.