Inteligentni sistemi **GRUPA I**

Prezime Ime: _____ Broj indeksa: _____

R uputstvo:

- U prilogu uz ovaj zadatak dat je materijal koji obuhvata dataset i cheatsheet. R projekat kreirati u pomenutom direktorijumu na destopu i nazvati ga RProjekat
- Zadatak se priznaje samo ako je kreiran PROJEKAT, inače se ne pregleda. Nije dovoljno kreirati samo R script fajl. R script fajl ne nazivati isto kao ceo projekat!
- Folder u kojem se nalazi kreirani projekat mora da sadrži SVE elemente koji su neophodni da se na nekom DRUGOM računaru (na kojem postoje instalirani R i RStudio) projekat otvori, da se u okviru otvorenog projekta otvori odgovarajući R script, kao i da se taj R script izvršava bez dodatnih intervencija

---

Zadatak (KMeans): Grupisanje pesama

U fajlu ***spotify.csv*** dati su podaci o najpopularnijim pesmama na Spotify platformi 2023. godine. Varijable dataset-a su:

- **track_name:** naziv pesme
- **released_year:** godina izdanja
- **in_spotify_playlists:** na koliko playlista na spotify platformi se nalzi pesma
- **in_spotify_charts:** pozicije pesme na spotify rang listi
- **streams:** broj puštanja pesme
- **in_apple_playlists:** na koliko playlista na appleplatformi se nalzi pesma
- **in_apple_charts:** pozicije pesme na applerang listi
- **bpm:** mera brzine pesme (beats per minute)
- **mode:** karakter pesme (minor / major)
- **energy:** percipirani energetski nivo pesme
- **instrumentalness:** nivo instrumentalnosti
- **liveness:** nivo prisustva elemenata uživo izvođenja
- **speechiness:** Skalirana količina izgovorenih reči u pesmi

---

Potrebno je uraditi sledeće:

1. Napraviti podskup podataka koji ne sadrži pesme čija je skalirana količina izgovorenih reči **(speechiness)** preko 60. Za sve naredne zahteve zadatka koristiti ovako dobijen dataset. **(1 poen)**
2. Odabrati atribute koji će biti uključeni u **KMeans model** i **navesti razlog** za njihov odabir ili neodabir. **(6 poena)**
3. U dataset-u dobijenom na osnovu prethodna dva zahteva proveriti da li postoje nedostajuće vrednosti (NA, “-”, ” ”, ili ””) i ako je moguće, zameniti ih adekvatnim vrednostima. Prokomentarisati postupak zamene vrednosti, odnosno zašto je određeni metod zamene vrednosti odabran. Tako dobijeni dataset eventualno **dodatno obraditi** da bi bio pogodan za primenu metode KMeans. **(8 poena)**
4. Primenom **Elbow metode** utvrditi najbolju vrednost za broj klastera **(k) (8 poena)**
5. Izvršiti klasterizaciju za izabranu (tj. utvrđenu najbolju) vrednost za **k. (2 poena)**
6. **Interpretirati**[1] dobijene klastere (grupisane pesme) na osnovu: broja pesama po klasteru, centara klastera, disperzije od centra. **(10 poena)**

Napomena: Varijable koje **očigledno** nema smisla da ostaviti u modelu **mogu izbaciti bez analiziranja** (npr. Izbacivanje ID-a) - potrebno je samo napisati komentar zasto se ta kolona izbacuje.

[1] Značenje tog izraza je: ***ne*** navoditi u komentarima samo brojke koje se dobiju, već protumačiti šta one ***znače*** u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi.