Inteligentni sistemi

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 desktopu 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 (KNN): Predikcija broja slušanja 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. Kreirati novu varijablu na osnovu vrednosti varijable **streams**. Varijablu nazvati **HighStreams**. Varijabla ima dve moguće vrednosti: *yes* (za vrednost **streams** iznad trećeg kvartila), i *no* (za sve ostale vrednosti). Pozitivna klasa je ***yes***. (**1 poen**)
2. Dataset treba eventualno dodatno pripremiti da bi bio pogodan za predviđanje vrednosti izlazne varijable **HighStreams** primenom ***kNN*** metode. U tom cilju, proceniti koje atribute je potrebno uključiti u model zasnovan na **kNN** algoritmu. **Obavezno** navesti u **komentaru** zašto su baš ti atributi uključeni u model. Takođe, ukoliko se neki atribut izostavi iz modela, **obrazložiti** zašto je izostavljen (**8 poena**).
3. U dataset-u dobijenom na osnovu prethodnog 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 **kNN**. (**7.5 poena**)
4. Primenom kros validacije sa 10 iteracija (10-fold cross-validation) odrediti najbolju vrednost za parametar K. (**5**.**5 poena**)
5. Kreirati klasifikacioni model na osnovu izabrane vrednosti za K. (**1 poena**)
6. Za kreirani model:
    - kreirati i interpretirati matricu konfuzije **u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi**[1] (**4 poena**)
    - navesti i objasniti 4 metrike koje se najčešće koriste za procenu klasifikatora, (**2 poena**)
    - izračunati i protumačiti vrednosti evaluacionih metrika **u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi**[2]. (**6 poena**)

Napomena: Varijable koje **očigledno** nema smisla koristiti za predviđanje izlazne se **mogu izbaciti bez analiziranja** (npr. Izbacivanje ID-a pri predikciji plate neke osobe) - potrebno je samo napisati komentar zasto se ta kolona izbacuje.

Svaka varijabla čiji odnos sa **izlaznom nije očigledan na prvi pogled**, mora se na neki način uporediti sa izlaznom, i na osnovu toga doneti zaključak da li se koristi u modelu li ne, sa propratnim komentarima.

---

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

[2] 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.