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 (NB): Klasifikacija jabuka

U fajlu *apples.csv* su dati podaci o jabukama uzgajanim na različitim kontinentima. Varijable dataset-a su:

- **A_id:** Identifikacioni broj jabuke.
- **Continent:** Kontinent na kojem je jabuka uzgajana.
- **Code:** Jedinstveni kod jabuke.
- **Size:** Veličina jabuke.
- **Weight:** Težina jabuke.
- **Sweetness:** Nivo slatkoće jabuke.
- **Crunchiness:** Hrskavost jabuke.
- **Juiciness:** Sokovitost jabuke.
- **Ripeness:** Zrelost jabuke.
- **Acidity:** Kiselost jabuke.
- **Quality:** Kvalitet jabuke

Potrebno je uraditi sledeće:

1. Kreirati novu varijablu na osnovu vrednosti varijable **Quality**. Varijablu nazvati **IsGood**. Varijabla ima dve moguće vrednosti: ***yes*** (za vrednost **Quality** = Good), 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 **IsGood** primenom **Naive Bayes** algoritma. U tom cilju, proceniti koje atribute je potrebno uključiti u model. **Obavezno** navesti u **komentaru** zašto su baš ti atributi uključeni u model. Takođe, ukoliko se neki atribut izostavi iz modela, **obrazložiti** zašto je izostavljen (**6 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 oblik zamene vrednosti odabran. Tako dobijeni dataset eventualno dodatno obraditi da bi bio pogodan za primenu metode **Naive Bayes**. (**5 poena**)
4. Kreirati model koji predviđa vrednost varijable **IsGood.** (**1.5 poena**)
5. Za kreirani model:
    - kreirati i interpretirati matricu konfuzije **u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi**[1] (**3.5 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]. (**8 poena**)
6. Pronaći optimalni prag klasifikacije (eng. probability threshold) kojim se maksimizuje suma specificity i sensitivity metrika. Za dobijeni threshold: (**8 poena**)
    - kreirati novu matricu konfuzije i izračunati evaluacione metrike,
    - **protumačiti rezultate u odnosu na rezultate sa default vrednošću praga klasifikacije**.

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.