ZADATAK: Predikcija odliva klijenata banaka

U fajlu *bank.csv* su dati podaci o **klijentima** evropskih banaka. Varijable dataset-a su:

- **Surname:** prezime
- **CreditScore:** kreditna sposobnost
- **Geography:** državljanstvo
- **Gender:** pol
- **Age :** godine
- **Tenure:** koliko dugo klijent ima nalog u banci (godine)
- **Balance:** stanje na računu
- **NumOfProducts:** broj proizvoda kupljenih korišćenjem usluga banke
- **HasCrCard:** da li klijent ima kreditnu karticu (0-ne, 1-da)
- **IsActiveMember:** da li klijent aktivno koristi usluge banke (0-ne, 1-da)
- **EstimatedSalary:** procenjena plata
- **Exited:** da li je klijent napustio banku (0-ne, 1-da)
- **Satisfaction.Score:** ocena zadovoljstva bankom
- **Card.Type:** tip kartice (premium program)
- **Point.Earned:** bodovi stečeni korišćenjem kartice

Potrebno je uraditi sledeće:

1. Kreirati novu varijablu na osnovu vrednosti varijable **Exited**. Varijablu nazvati **Stayed**. Varijabla ima dve moguće vrednosti: *1* (Yes; za vrednosti 0 varijable **Exited**), i *0* (No; za sve ostale vrednosti). Pozitivna klasa je ***1***.
2. Napraviti podskup podataka koji **ne** sadrži klijente koji imaju više od 87 godina.
3. Proceniti koje atribute je potrebno uključiti u model **stabla odlučivanja.** Obavezno navesti u komentaru zašto su baš ti atributi uključeni u model. Takođe, ukoliko se neki atribut izostavi, obrazložiti zašto je izostavljen.
4. U dataset-u dobijenom na osnovu prethodna tri 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 **stabla odlučivanja**.
5. Primenom kros validacije sa 10 iteracija (10-fold cross-validation) odrediti najbolju vrednost parametra ***ccp_alpha*** za predviđanje vrednosti **Stayed**.
6. Kreirati klasifikacioni model na osnovu određene optimalne vrednosti za ***ccp_alpha***.
7. Za kreirani model:
    - kreirati i interpretirati matricu konfuzije **u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi**[1]
    - navesti i objasniti 4 metrike koje se najčešće koriste za procenu klasifikatora,
    - izračunati i protumačiti vrednosti evaluacionih metrika **u kontekstu problema koji se rešava u zadatku i dataset-a koji se koristi**[2].

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.