Prvi deo projektnog zadatka – analiza baze podataka

U nastavku je dato 16 pitanja. Potrebno je napisati kratke odgovore na svako od 16 pitanja i predati putem moodle stranice u formi pdf-a do 15.12.2023. Ako imate bilo kakvih nedoumica, obratite se asistentima na času ili putem mail-a.

1. Definisati u 2-3 rečenice problem koji će se u projektu rešavati. (primer: Rešavaće se problem detekcije karcinoma dojke na osnovu analiza iz krvi. U pitanju je klasifikacioni problem sa 2 klase.)
2. Koliko ima uzoraka u bazi?
3. Jednom rečenicom objasniti šta predstavlja jedan uzorak u konkretnoj bazi.
4. Koliko ima obeležja u bazi?
5. Navesti sva obeležja (jasnim imenom na srpskom ili opisno, nebitan je naziv u samoj bazi).
6. Koliko ima numeričkih obeležja?
7. Ako ima kategoričkih obeležja, navesti koje od njih ima najmanji broj kategorija i koje su, i navesti ono koje ima najveći broj kategorija i koliko ih je.
8. Ako se rešava regresioni problem: navesti opseg, sr.vr. i medijanu obeležja koje će se predvidjati.
Ako se rešava klasifikacioni problem: navesti procentualno koliko ima uzoraka u svakoj od klasa.
9. Da li postoje obeležja u bazi koja smatraš da treba izbaciti iz baze? Koja su to i zašto smatraš da ih treba izbaciti?
10. Da li u bazi ima nedostajućih vrednosti? Ako ima, navesti za svako od obeležja koliko vrednosti mu procentualno nedostaje?
11. Da li ima nevalidnih vrednosti u bazi? Ako ima, navesti za svako od obeležja koje su vrednosti nevalidne i zašto se smatraju nevalidnim.
12. Ako ima nedostajućih i/ili nevalidnih vrednosti u bazi, za svako od obeležja navesti kako će problem biti rešen.
13. Kada je završeno izbacivanje, dopuna, i drugo, navesti koliko je u sređenoj bazi ostalo uzoraka, a koliko obeležja.
14. Da li neka od obeležja sadrže autlajere? Nabvesti koja obeležja ih sadrže.
15. Da li postoje parovi obeležja korelisani više od 0.7? Navesti takve parove obeležja.
16. Ako se rešava klasifikacioni problem: iscrtavanjem histograma reći koje se obeležje izdvaja kao najdiskriminatornije (najbolje razdvaja klase)?
Ako se rešava regresioni problem: utvrditi koliko je odstupanje raspodele varijable koja se predviđa od normalne raspodele dobijene korišćenjem uzoračke sr.vr. i standardne devijacije (asimetričnost i spljoštenost)?