OZP – Pitanja za drugi projektni zadatak

**Klasterovanje:**

1. Šta je to klastering i čemu služi? (definicija i primer primene)
2. Razlika između segmentacije i klasteringa.
3. Kako je moguće vizualizovati klaster modele?
4. Objasniti K-means algoritam (koraci).
5. Prednosti i mane K-means algoritma.
6. Objasniti euklidsko odstojanje (kako se računa).
7. Objasniti city-blok odstojanje (kako se računa).
8. Kako se određuje broj klastera kod K-means algoritma.
9. Objasniti K-mens++ inicijalizaciju.
10. Šta su to outlajeri?
11. Šta je to normalizacija?
12. Objasniti Box-plot metodu za identifikaciju i uklanjanje outlajera.
13. Objasniti Z-score metodu za identifikaciju i uklanjanje outlajera.
14. Objasniti proces evaluacije klaster modela.
15. Objasniti Silhouette indeks.
16. Objasniti SSE (kompaktnost).
17. Objasniti „Lakat“ metodu (kako se sprovodi i čemu služi).
18. Kako se može vizualno predstaviti K-means model?
19. Objasniti Silhouette score.
20. Objasniti Hijerarhijsko aglomerativno klasterovanje.
21. Navesti i objasniti metode povezivanja (*linkage* metode)
22. Šta je to dendrogram.
23. Kako se dendrogram koristi za određivanje broja klastera?

**Klasifikacija:**

1. Šta je to klasifikacija i dati barem dva primera primene (koji nisu rađeni na nastavi).
2. Razlika između klasifikacije i klasteringa.
3. Navedite i kratko objasnite različite modele/algoritme za klasifikaciju.
4. Šta je to granica odlučivanja i kompleksnost modela, i u kakvoj su vezi?
5. U kakvoj su vezi kompleksnost modela i sklonost modela ka pretreniranju?
6. Kako se kontroliše kompleksnost modela stabla odlučivanja?
7. Objasniti šta je to pretreniranje (*eng*. *overfitting*).
8. Objasniti šte je to pristrasnost (eng. bias, *underfitting*).
9. Na koji način se može prepoznati pretreniran model?
10. Na koji način se može sprečiti pretreniranje modela?
11. Šta je to greška generalizacije?
12. Zbog čega se iz podataka izdvaja test uzorak?
13. Šta je nedostatak kada se izdvoji samo jedan test uzorak i kako se rešava?
14. Šta je to kros-validacija i kako se sprovodi?
15. Objasniti kada treba koristiti veliko k (broj podgrupa) u kros-validaciji.
16. Šta je to matrica konfuzije i kako se tumači?

17. Navesti barem 3 mere za evaluaciju modela klasifikacije.
18. Preciznost (eng. *Precision*) – objasniti kako se računa i čemu služi.
19. Odziv (eng. *Recall*) – objasniti kako se računa i čemu služi.
20. F-mera (eng. *F-score*) – objasniti kako se računa i čemu služi.
21. Šta treba uraditi ukoliko je greška na test skupu mnogo veća od greške na trening skupu?
22. Šta treba uraditi ukoliko je greška na trening skupu prevelika da bi se model koristio u praksi?
23. Objasniti razliku između TP (eng. *True positive*) i FN (eng. *False Negative*) slučajeva.
24. Objasniti razliku između FN (eng. *False negative*) i FP (eng. *False positive*) slučajeva.
25. Ako se rešava problem odobravanja kredita, koja mera performansi klasifikacionog modela je adekvatnija?
26. Ako se rešava problem predviđanja koji lek davati pacijentu, koja mera performansi je adekvatna?
27. Ako se pravi model koji klasifikuje decu u buduće delikvente ili ne-delikvente, radi dodatne edukacije i pomoći države, koja mera performanse je adekvatna?
28. Kako se pomeranjem granica odluke (threshold) može unaprediti izgrađeni model?
29. Šta je ROC kriva i mera AUC?

**Manipulacija podataka u jeziku *Python*.**

Na odbrani može biti dat zahtev pojedincima da nad podacima izvrše određenu manipulaciju: selektuju određene redove/kolone po indeksu, selektuju redove/kolone koji zadovoljavaju određeni uslov, naprave pivot tabelu, izbace neku kolonu, prikažu tipove podataka za svaku kolonu, učitaju/sačuvaju podatke, itd.