1. Navedite dva pristupa smanjenju broja obeležja. Po čemu se ovi pristupi razlikuju (iz perspektive rezultujućeg smanjenog skupa obeležja)?

2. PCA je tehnika za _____.

3. Objasnite i skicirajte osnovnu ideju PCA (ne zaboravite da označite ose na graficima).

4. U kakvom su odnosu obeležja koja proizvede PCA algoritam sa originalnim obeležjima?

5. Ako je p_1 prva glavna komponenta, a p_2 druga, koji od sledećih iskaza su tačni:
   1) Varijansa po p_2 je veća od varijanse po p_1.
   2) p_1 je normalna na p_2.
   3) p_1 je paralelna sa p_2.
   4) Varijansa po p_1 je veća od varijanse po p_2.

6. Da li je potrebno pretprocesirati skup podataka pre primene PCA? Ako jeste, napišite kako i zbog čega.

7. Šta je pravilan postupak za normalizaciju obeležja (1, 2 ili svejedno):
   1) Prvo, za svako obeležje d odredimo srednju vrednost μ_d i standardnu devijaciju σ_d. Normalizujemo sva obeležja primenom formule $x_d = \frac{(x_d - \mu_d)}{\sigma_d}$. Nakon toga, podelimo skup podataka na trening i test skup. Treniramo model na trening skupu, a potom ga evaluiramo na test skupu.
   2) Podelimo skup podataka na trening i test skup. Za svako obeležje, koristeći isključivo trening skup, odredimo srednju vrednost srednju vrednost μ_d i standardnu devijaciju σ_d. Potom, normalizujemo sva obeležja trening skupa i sva obeležja test skupa primenom iste formule $x_d = \frac{(x_d - \mu_d)}{\sigma_d}$. . Treniramo model na trening skupu, a potom ga evaluiramo na test skupu.

   Obrazložite svoj odgovor.

8. PCA je nadgledana/nenadgledana metoda (zaokružite tačno).

9. Kako odabrati K (novi broj dimenzija) kod PCA metode?
   1) Ako je cilj vizuelizacija:
   2) Ako je cilj da ubrzamo izvršavanje algoritma (uz izbegavanje prevelikog gubitka u performansama):

10. Da li se PCA može primeniti za prevenciju overfittinga?

11. PCA radi bolje ako: (1) postoji linearna struktura u podacima, (2) podaci leže na zakrivljenoj površini, a ne na ravnoj, (3) obeležja su skalirana na iste opsege:
    1) Važe (1) i (2)
    2) Važe (2) i (3)
    3) Važe (1) i (3)
    4) Važe (1), (2) i (3).

12. Tačno ili netačno:
    1) PCA je podložan upadanju u lokalni optimum. Ispitivanje više nasumičnih inicijalizacija može da pomogne.
    2) PCA je deterministički algoritam.
    3) Sve glavne komponente dobijene pomoću PCA su međusobno ortogonalne.
    4) Čak i ako se ulazna obeležja kreću u veoma sličnim opsezima, treba da centriramo podatke pre pokretanja PCA.
    5) Ukoliko imamo n-dimenzione podatke, ima smisla da pokrećemo PCA sa $k \leq n$, gde je k željeni broj dimenzija. Konkretno, pokretanje PCA sa $k = n$ je moguće ali nije od koristi, dok $k > n$ nema smisla.
    6) Treba da uklonimo visoko korelirana obeležja pre primene PCA.
    7) Recimo da iskoristimo PCA da projektujemo d-dimenzione tačke u j-dimenzioni prostor. Zatim, ponovo pokrenemo PCA da projektujemo te tačke iz j-dimenzionog prostora u k-dimenzioni prostor ($d > j > k$). Dobićemo isti rezultat kao da smo iskoristili PCA da direktno projektujemo d-dimenzione tačke u k-dimenzioni prostor.
    8) Recimo da dodamo 1 na kraj vektorske reprezentacije svake instance u skupu podataka. Ovo neće promeniti rezultate PCA (osim što će korisne glavne komponente imati 0 na kraju i imaćemo jednu dodatnu beskorisnu komponentu sa sopstvenom vrednošću 0).
    9) I PCA i LDA su linearne transformacije.
    10) LDA je nadgledan metod, dok je PCA nenadgledan.
    11) PCA maksimizuje varijansu u podacima, dok LDA maksimizuje separaciju između različitih klasa.
13. Koje od sledećih su preporučene primene za PCA?
    1) Vizuelizacija podataka: da uzmemo 2D podatke i pronađemo drugi način da ih plotujemo u 2D (koristimo $k = 2$).
    2) Kao zamena (ili alternativa) linearnoj regresiji: u dosta primena, PCA i linearna regresija daju slične rezultate.
    3) Kompresija podataka: pre obučavanja modela nadgledanog učenja, smanjimo dimenzionalnost ulaznih podataka kako bismo ubrzali učenje.
    4) Kompresija podataka: smanjimo dimenzionalnost podataka kako bi zauzimali manje prostora na disku/u memoriji.
14. Šta će se desiti ako su sopstvene vrednosti približno jednake?
    1) PCA će imati odlične performanse
    2) PCA će raditi loše (nećemo moći da odaberemo glavne komponente)
    3) Ne možemo reći.
15. Kako možemo evaluirati performanse algoritma za redukciju dimenzionalnosti?

16. Šta od sledećih mogu da budu prve dve glavne komponente nakon primene PCA?
    1) (0.5, 0.5, 0.5, 0.5) i (0.71, 0.71, 0, 0)
    2) (0.5, 0.5, 0.5, 0.5) i (0, 0, -0.71, -0.71)
    3) (0.5, 0.5, 0.5, 0.5) i (0.5, 0.5, -0.5, -0.5)
    4) (0.5, 0.5, 0.5, 0.5) i (-0.5, -0.5, 0.5, 0.5)
17. Šta je tačno kada imamo podatke projektovane u prostor niže dimenzionalnosti dobijen pomoću PCA?
    1) Nova obeležja su i dalje interpretabilna.
    2) Nova obeležja gube interpretabilnost
    3) Nova obeležja sigurno sadrže sve informacije iz originalnog skupa podataka
    4) Nova obeležja ne moraju da sadrže sve informacije iz originalnog skupa podataka.