1. Kada biste, u cilju uklanjanja nedostajućih vrednosti, primenili pročišćavanje podataka (uklanjanje obeležja ili opservacija iz skupa podataka)?

2. Glavni nedostatak strategije uklanjanja nedostajućih vrednosti je _____.

3. Ukoliko ne želimo da uklonimo nedostajuće vrednosti, možemo ih proceniti. Za kategorijska obeležja tipično koristimo _____, a za numerička _____. Postoje i naprednije tehnike kao što su _____. Glavni nedostatak ove strategije je _____.

4. *Naive Bayes* (NB) model je pogodan za _____(male/velike) skupove podataka jer je skloniji da pati od nego od _____ (sistematsko odstupanje/varijansa).

5. Tačno ili netačno (iskazi se odnose na NB model):
   a. Sva obeležja su jednako važna.
   b. Obeležja su statistički zavisna jedna od drugih za zadatu klasu.
   c. Obeležja su statistički nezavisna jedna od drugih za zadatu klasu.
   d. Obeležja mogu biti kategorička ili numerička.

6. Formula data Bajesovom teoremom je:

7. U formuli, aposteriorna verovatnoća je _____, a apriorna _____. Dokaz je _____.

8. Problem:

   Želimo da odredimo da li pacijent ima određenu formu raka. Znamo da svega 0.8% ljudi na svetu ima ovu formu raka. Postoji test krvi koji nam vraća POS i NEG rezultat. Ako osoba nema rak, dobiće NEG rezultat u 97% slučajeva. Ako osoba ima rak, dobiće POS rezultat u 98% slučajeva.

   Apriorna verovatnoća klase „nema rak“ je _____.

9. Kako interpretirate „dokaz“ u Bajesovoj teoremi?

10. Nastavak na problem iz zadatka 9: Zašto nam je za određivanje $P(cancer|POS)$ potrebna Bajesova teorema? Odnosno, zašto ovu verovatnoću procenjujemo primenom formule $P(\text{cancer}|\text{POS}) = \frac{P(\text{POS}|\text{cancer})\text{P}(\text{cancer})}{P(\text{POS})}$, a ne procenjujemo $P(cancer|POS)$ direktno iz skupa podataka?

11. Zašto se NB klasifikator zove „naivan“?

12. Zašto uvodimo „naivnu“ pretpostavku?

13. Kada računamo verovatnoću $P(x_d|y = c) = \frac{N_{x_d,c}}{N_c}$ u NB modelu, na koju grešku možemo naići i kako je možemo rešiti?

14. Dva načina da primenimo NB u slučaju kontinualnih obeležja su _____ i _____.

15. Tačno ili netačno:

    a. NB je generativni model.
    b. NB je robustan na irelevantna obeležja.
    c. NB ima dobre performanse, čak i kada je narušena pretpostavka o uslovnoj nezavisnosti obeležja date klase.
    d. Logistička regresija je generativni model.
    e. NB je nelinearni klasifikator.
    f. Generativni modeli direktno uče $P(y|x)$.
    g. Generativni modeli uvode slabije pretpostavke o podacima od diskriminativnih modela.
    h. Generativnim modelima treba manje podataka za trening od diskriminativnih modela.
    i. Generativni modeli mogu da rade sa nedostajućim vrednostima, a diskriminativni ne.
    j. Obučenim generativnim modelom možemo generisati primere skupa podataka.

16. Koju pretpostavku uvodimo kod NB sa Gausovim kernelom? Kako se obezbeđujemo da je ispoštovana?