ZADATAK 1:
Q: Koja klasa ima najviše pogrešnih pozitivnih klasifikacija (engl. false positives), a koja najviše pogrešnih negativnih klasifikacija (engl. false negatives)?
A: Kod mene klasa 0 ima najvise (16) false positive, a klasa 1 najvise (36) false negative klasifikacija.

Q: Što predstavljaju vrijednosti na dijagonali ove matrice?
A: Broj ispravno klasificiranih primjera.

Q: Koja je razlika između mikro- i makro-procjene?
A: Mikro prosjek računamo izravno na tablici zabune (pribrajamo TP,FP,FN,TN za svaku klasu te potom računamo mjeru), a za makro prosjek računamo K vrijednosti mjera na tablicama zabune 2x2 te zatim uprosječimo. Makro procjena je tipično manja od mikro procjene.

ZADATAK 2:
Q: Zašto nam treba još jedna razina k-struke provjere?
A: Zato što osim vrednovanja klasifikatora istovremeno radimo i optimizaciju hiperparametara modela - jedna k-struka provjera nam je potrebna za utvrđivanje optimalnih hiperparametara, a druga za vrednovanje.

ZADATAK 3:
Q: Koju hipotezu H0 i alternativnu hipotezu H1 testiramo ovim testom?
A: H0 - ne postoji razlika u srednjim vrijednostima dviju populacija, mi1 = mi2. H1 - postoji razlika u srednjim vrijednostima dviju populacija, mi1 != mi2.

Q: Ako odaberemo razinu značajnosti α=0.05, hoćemo li odbaciti hipotezu H0?
A: Nećemo - dobivena p-vrijednost je 0.3996, a ona je veća od razine značajnosti alfa te se hipoteza ne odbacuje.

Q: Koja pretpostavka na vjerojatnosnu razdiobu primjera je napravljena u gornjem testu? Je li ona opravdana?
A: Pretpostavka da se primjeri pokoravaju studentovoj razdiobi. Ta je pretpostavka opravdana zato što su naši podaci srednje vrijednosti pogreške, a distribucija uzorkovanja srednje vrijednosti općenito teži k normalnoj distribuciji (zbog centralnog graničnog teorema). Studentova razdioba za velik uzorak jako sliči normalnoj razdiobi te je stoga njeno korištenje opravdano.

Q: Koji je model u konačnici bolji i je li ta prednost značajna uz α=0.05?
A: (u mom slučaju) Iako je dobiveni prosjek generalizacijske pogreške SVMa manji od prosjeka generalizacijske pogreške logističke regresije, odgovor je da NE ZNAMO koji je model bolji jer nismo odbacili hipotezu uz razinu značajnosti alfa=0.05.

ZADATAK 4:
Q: Koju biste vrijednost hiperparametra KK izabrali na temelju ovog

grafa? Zašto?
A: Vrijednost K=3, jer se na tom mjestu dogodio pregib (dogodila se nagla promjena kriterijske funkcije algoritma k-srednjih vrijednosti).

Q: Je li ova metoda robusna?
A: Ne nužno. Ako imamo dosta snažnih outliera, oni mogu znatno povećati iznos kriterijske funkcije te će zahtijevati dodatne grupe (veći parametar K), iako u stvarnosti ne postoji potreba za tim.

Q: Znate li još koji način kako izabrati vrijednost hiperparametra K?
A: skripta, str. 123-124

Q: Možemo li izabrati onaj K koji minimizira pogrešku J? Objasnite.
A: Ne možemo. Pogrešku J minimizirat će K koji je jednak broju primjera, K=N, tj. tada će svaki primjer biti centroid za sebe i pogreška J će biti J=0. Mi tražimo onaj K koji ima najbolju sposobnost generalizacije, a preveliki K će, iako smanjuje kriterij J, dovesti do prenaučenosti modela.

Q: Kako biste svojim riječima opisali što mjeri Randov indeks?
A: (ovo je sjebano objasniti, ja i dalje ne kužim baš) Randov indeks mjeri sličnost između dva skupa grupiranja podataka. Uzima parove podataka iz dva skupa grupiranja podataka te penalizira jednako označene parove u različitim grupama i različito označene parove u istim grupama.

Q: Koji su glavni problemi ove metrike?
A: Glavni je problem to da očekivanje Randovog indexa dvaju random grupiranja nema konstantnu vrijednost. To znači da u biti ne znamo što je visok, a što nizak Randov index jer nemamo referentnu vrijednost s kojom možemo usporediti. Npr. kod klasifikacije, ako napravimo klasifikaciju 0-1 na random način, očekivana pogreška je 50% i sve manje od toga je dobro. Kod Randovog indeksa ne znamo očekivanu pogrešku random grupiranja pa je teško samo na temelju njega reći radi li naše grupiranje dobro.

ZADATAK 5:
Q: Što se dogodilo? Koja je pretpostavka algoritma k-srednjih vrijednosti ovdje narušena?
A: Dio primjera jedne grupe bio je bliži centroidu druge grupe pa se tako (krivo) grupirao. Narušena je pretpostavka da svaki primjer pripada grupi čiji mu je centroid najbliži (jer jedna grupa ima mnogo veću standardnu devijaciju od druge, neki od njenih primjera će biti blizu centra druge grupe te se krivo grupirati).
Q: Što biste morali osigurati kako bi algoritam pronašao ispravne grupe?
A: Nemam pojma. Ili osigurati da su standardne devijacije dvaju grupa jednake (što nema smisla kad u praksi ne znamo kako su modelirani podaci - zato i radimo grupiranje) ili koristimo drugi algoritam.

Q: Što se dogodilo? Koja je pretpostavka algoritma k-srednjih

vrijednosti ovdje narušena?
A: Kako su dva kruga jedan unutar drugog (linearno neodvojivi podaci, u stvarnosti su centroidi obje grupe isti), algoritam k-srednjih vrijednosti ne može točno grupirati. Narušena je pretpostavka da su centroidi dvaju grupa različiti.
Q: Što biste morali osigurati kako bi algoritam pronašao ispravne grupe?
A: Osigurati da su centroidi dvaju grupa različiti (nemoguće za ovakve, kružne podatke) ili promijeniti algoritam.

Q: Što se dogodilo? Koja je pretpostavka algoritma k-srednjih vrijednosti ovdje narušena?
A: Zelena grupa ima jako puno podataka, a roza grupa malo podataka. Centroid zelene grupe je određen i dosta je fiksan, teško ga je pomaknuti - iako je neki podatak daleko od centra, on ne utječe puno na sam centroid zbog mnogo podataka. S druge strane, centroid roze grupe je lako pomaknuti jer imamo malo podataka. Neki dijelovi zelene grupe se u startu grupiraju pod rozu grupu, a to pomiče centroid roze grupe bliže zelenoj te onda još više zelenih podataka upada u rozu grupu i centroid roze grupe se još više približava zelenoj te se to praktički rekurzivno ponavlja - roza grupa "izjeda" zelenu te je na kraju dosta podataka krivo grupirano. Narušena je pretpostavka da grupe imaju otprilike jednak broj članova.
Q: Što biste morali osigurati kako bi algoritam pronašao ispravne grupe?
A: Da grupe imaju podjednak broj članova ili koristiti drugi algoritam.

Q: Uspjeva li GMM riješiti "probleme" koje ima algoritam k-srednjih vrijednosti? Zašto?
A: Prvi primjer da i to savršeno, jer su naši podaci u tom slučaju modelirani Gaussovim distribucijama pa će GMM naći parametre koji točno grupiraju. Drugi primjer ne, jer GMM isto kao i K-means pretpostavlja da podaci imaju različite centre. Treći primjer djelomično - kako GMM modelira Gaussove distribucije, on je rozu grupu modelirao s točnim središtem, ali previsokom varijancom pa je dio zelene grupe s lijeve strane krivo pripao rozoj grupi.