Istraživanje podataka 1 - zadaci

1. Python
   - Generisati skup podataka korišćenjem date funkcije *generate_data*.
   - Koliko ima različitih klasa?
   - Ako je potrebno, podeliti skup podataka na deo za obučavanje i deo za testiranje modela.
   - Da li je skup balansiran? Ako je potrebno, primeniti SMOTE algoritam.
   - Primeniti SVM algoritam sa linearnim i rbf kernelom za klasifikaciju.
   - Uz pomoć PCA algoritma smanjiti dimenzionalnost na 2. Koji udeo varijanse je očuvan? Da li je to očekivano s obzirom na to kako su podaci generisani?
   - Nacrtati grafik tako transformisanog skupa – obojiti instance u skladu sa klasom kojom pripadaju, a posebnom bojom označiti potporne vektore.
   - Proceniti kvalitet dobijenih modela koristeći matricu konfuzije i tačnost (accuracy) modela. Da li je tačnost adekvatna metrika za ovaj skup podataka?
     Da li su dobijeni rezultati očekivani s obzirom na svojstva korišćenih kernela i izgled skupa podataka?

2. Python
   - Učitati skup podataka *20newsgroups.*
   - Ograničiti broj instanci na 200, vodeći računa o raspodeli klasa.
   - Koliko ima različitih klasa? Ako je potrebno, podeliti skup podataka na deo za obučavanje i deo za testiranje modela.
   - Pretvoriti tekstualne podatke u TF-IDF matricu.
   - Da li je potrebno dodatno pretprocesiranje podataka? Objasniti odgovor i ako je potrebno primeniti metode pretprocesiranja.
   - Primeniti K-means algoritam za klasterovanje sa 20 klastera.
   - Primeniti hijerarhijsko klasterovanje sa 20 klastera.
   - Ispisati silueta koeficijente za pronađene klastere. Da li su njihove vrednosti dobre? Koja metoda klasterovanja nam daje bolju vrednost? Kolika bi bila idealna vrednost silueta koeficijenta? Da li je silueta koeficijent adekvatna metrika za ovaj skup? Ako ne, zbog čega?
   - Primeniti PCA proceduru sa dve komponente nad ulaznim podacima. Koliki procenat varijanse se ovime održao? Nacrtati grafik transformisanog skupa u dve dimenzije i instance obojiti na osnovu toga kom klasteru pripadaju (jedan grafik za K-means, a jedan grafik za hijerarhijsko klasterovanje).
   - Uporediti klastere dobijene pomoću ove dve metode sa originalnim klasama datim u skupu podataka. Koliko se dobro podudaraju?

3. SPSS
   - Učitati skup podataka *balloons.csv.*
   - Da li ima nedostajućih vrednosti? Ako ih ima, izbaciti instance koje ih sadrže.
   - Primeniti algoritam C5.0 i zadati da je minimalan broj instanci koji mora da bude u dete-čvoru 5. Dobijeni model nazvati *model1*.
   - Da li je potrebno dodatno pretprocesiranje podataka? Objasniti odgovor i ako je potrebno primeniti metode pretprocesiranja.
   - Koji atributi su najznačajniji za pravljenje modela *model1*?
   - Podatke o dobijenom modelu (tačnost i matrice konfuzije na trening i test skupu) sačuvati u html datoteci.
   - Prodiskutovati kvalitet dobijenog modela.