

Istraživanje podataka 1

Januar 1 PS

Školska godina 2025/26

Uputstvo:

Na Desktop-u se nalazi folder sa nazivom `ip1_rok_miGGIII_ime_prezime` u kome se nalaze zadaci i skupovi podataka. Folder preimenovati tako da odgovara vašim podacima i sva rešenja zadataka čuvati u njemu, pri čemu su odgovarajući fajlovi imenovani sa *1.ipynb*, *2.ipynb* i *3.ipynb*.

1. Klasifikacija na skupu podataka `diabetes.csv`

- Prikazati osnovne deskriptivne statistike (prosek, standardna devijacija, kvartili).
- Ukoliko ima nedostajućih vrednosti, gde za kolone `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI` vrednost 0 označava nedostajuću vrednost, zameniti ih prosekom za tu kolonu.
- Odrediti elemente van granica (eng. *outliers*) korišćenjem standardne devijacije. Ukoliko postoje, zameniti ih graničnim vrednostima.
- Klasifikovati podatke (ciljna kolona je `Outcome`) korišćenjem *SVM* algoritma. Da li je potrebno podeliti podatke na trening i test skup? Da li je potrebno primeniti neki vid normalizacije podataka? Zašto?
- Prikazati matricu konfuzije, tačnost i *F1*. Oceniti kvalitet modela.
- Uporediti rezultate linearnog i kernelizovanog *SVM*-a.
- Da li je skup balansirani? Ukoliko nije, primeniti *SMOTE* i obučiti novo stablo odlučivanja na izbalansiranom skupu. Uporediti rezultate sa originalnim modelom.

2. Klasterovanje na skupu podataka `circles.csv`

- Primenom algoritma *DBSCAN* pronaći 2 klastera.
- Primenom algoritma sakupljajućeg hijerarhijskog klasterovanja pronaći 2 klastera. Koristiti euklidsko rastojanje.
- Uporediti rezultate klasterovanja u odnosu na tip veze (*min*, *max*, *avg*), kao i u odnosu na *DBSCAN*. Da li su rezultati očekivani? Zašto?
- Prikazati grafički rezultate koristeći *scatter plot*. Obojiti instance na osnovu klastera kom pripadaju.
- Da li je bilo potrebno podeliti podatke na trening i test skup? Da li je bilo potrebno dodatno pretprocesiranje podataka? Zašto?

3. Pravila pridruživanja na skupu podataka `market.csv`

- Primeniti algoritam apriori sa minimalnom podrškom od 20% i minimalnom pouzdanošću 60%.
- Koliko pravila je ukupno pronađeno?
- Koliko pravila je interesantno po *lift* meri?
- Koliko ima pravila u kojima je u zaključku sir (*Cheese*)? Izdvojiti sva takva u pravila u poseban model.
- Prikazati grafik gde su kolone čvorovi, a debljina grane između njih je proporcionalna broju zajedničkog pojavljivanja.
- Da li je bilo potrebno podeliti podatke na trening i test skup? Da li je bilo potrebno dodatno pretprocesiranje podataka? Zašto?