

**Istraživanje podataka 1**  
*Vanredni rok - Decembar*  
Školska godina 2024/2025

**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 `wine.csv`.

- Prikazati osnovne deskriptivne statistike (prosek, standardna devijacija, kvartili).
- Ukoliko ima nedostajućih vrednosti, izbaciti instance koje ih sadrže.
- Odrediti elemente van granica (eng. *outliers*) korišćenjem interkvartilnog raspona. Ukoliko postoje, zameniti ih graničnim vrednostima.
- Klasifikovati podatke (ciljna kolona je **target**) korišćenjem stabla odlučivanja.
- Da li je potrebno podeliti podatke na trening i test skup? Da li je potrebno primeniti neki vid normalizacije podataka?
- Prikazati matricu konfuzije, tačnost i *F1*. Oceniti kvalitet modela.
- Obučiti model slučajne šume i uporediti rezultate.
- Primenom *PCA* smanjiti dimenzionalnost podataka na 2. Koji udeo početne varijanse podataka je očuvan?
- Tako transformisan skup prikazati grafički korišćenjem *scatter plot*a. Obojiti instance na osnovu predviđene klase.

2. Klasterovanje na skupu podataka `moons.csv`.

- Primenom algoritma K-means pronaći 2 klastera.
- Prikazati grafički rezultate koristeći *scatter plot*. Obojiti instance na osnovu klastera kom pripadaju. Centroide označiti crnom bojom.
- 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 *K-means*. Da li su rezultati očekivani?
- 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. Klasifikacija na skupu podataka `20newsgroups.csv`.

- Klasifikovati podatke (ciljna kolona je **target**) korišćenjem naivnog Bajesovog algoritma i *TF* matrice.
- Da li je potrebno podeliti podatke na trening i test skup?
- Prikazati matricu konfuzije, tačnost i *F1*. Oceniti kvalitet modela.
- Formirati *TF-IDF* matricu, obučiti novi model naivnog Bajesa i uporediti rezultate.
- Izdvojiti 5 instanci koje su pogrešno klasifikovane i ispisati ih.
- Ukoliko je moguće, obučiti i linearni *SVM* model i uporediti rezultate.