# Mašinsko učenje 2024

**Zadatak 5**

# Sadržaj

# Zadatak 4 - Rekapitulacija

# Zadatak 4 - Rekapitulacija

- Procenat uspešnosti: **82%** (28/34).
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **24%**.
- Najbolji rezultati po terminima:

| Termin | Tim | Macro F1 |
|---|---|---|
| Ponedeljak - G4 | **tim1_24** | **0.43** |
| Utorak - G5 | **tim11_24** | **0.45** |
| Utorak - G3 | **Tim 10** | **0.46** |
| Četvrtak - G2 | **Ruzni kao passsss** | **0.47** |
| Petak - G1 | **tim21_24** | **0.46** |

# Zadatak 4 - Rekapitulacija

- Dobre stvari (na nivou generacije):
  - Pretprocesiranje
  - Propratni izveštaji.

- Stvari koje mogu biti bolje (na nivou generacije):
  - Uklanjanje *outlier*-a
  - Rad sa *LabelEncoder*-om.

# Zadatak 5

# Zadatak 5

- Klasterovanje:
  - Klasterovati države na osnovu njihovih karakteristika u klastere koji predstavljaju geografske regione (kolona **region**):
    - **europe**
    - **asia**
    - **africa**
    - **americas.**
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije **v mera** (eng. *v measure score*) veća od 0.13.
  - Zadatak se rešava upotrebom Modela Gausovih mešavina (eng. *Gaussian Mixture Model, GMM*), tj. algoritmom Očekivanje - maksimizacija (eng. *Expectation-maximization, EM*).
  - Rok za izradu zadatka je **02.06.2024. u 23:59h**.

# Zadatak 5

- Redukcija dimenzionalnosti:
  - Instalirane biblioteke za Zadatak 5:
    - **NumPy**
    - **Pandas**
    - **SciPy**
    - **scikit-learn**.

# Zadatak 5

- Sledeći i poslednji termin vežbi (odbrana Zadatka 5):

| Termin | Datum |
|---|---|
| Ponedeljak - G4 | **10.06.2024.** |
| Utorak - G5 | **04.06.2024.** |
| Utorak - G3 | **04.06.2024.** |
| Četvrtak - G2 | **06.06.2024.** |
| Petak - G1 | **07.06.2024.** |

# Zadatak 5

- Atributi:
  - **Year** - Godina za koju važe navedeni podaci
  - **Population** - Ukupan broj stanovnika
  - **GDP per Capita** - GDP prihod države u toj godini (izražen u $)
  - **Urban Population** - Broj stanovnika u urbanim podnebljima
  - **Life Expectancy** - Očekivana dužina životnog veka
  - **Surface Area** - Površina države
  - **Literacy Rate** - Procenat pismenih stanovnika.

# Zadatak 5

- Koncepti vezani za Zadatak 5:
  - Modeli Gausovih mešavina
  - *Expectation-Maximization* algoritam
  - Metrika.

# Zadatak 5

- Trening skup podataka sadrži nedostajuće vrednosti (u pitanju su prazne ćelije).
- Testni skup podataka **ne** sadrži nedostajuće vrednosti.

# Zadatak 5

- Modeli Gausovih mešavina:
  - Probabilistički modeli koji pretpostavljaju da su podaci predstavljeni kao zbir više Gausovih (normalnih) raspodela
  - Koriste se za probleme klasterizacije gde se podaci dele na klastere na osnovu verovatnoće pripadnosti svakoj Gausovoj komponenti
  - Modeli Gausovih mešavina u scikit-learn.

# Zadatak 5

- *Expectation-Maximization* algoritam:
  - Iterativna metoda za pronalaženje maksimalno verovatnih parametara u statističkim modelima
  - U scikit-learn biblioteci se ovaj algoritam koristi za estimaciju parametara GMM
  - Zadatak se **mora** rešiti upotrebom GMM, tj. EM algoritma.
  - Algoritam se može samostalno implementirati, a može se iskoristiti i implementacija scikit-learn biblioteke.

# Zadatak 5

- Metrika:
  - Kod ovog zadatka, evaluacija klasterovanja je zasnovana na poznavanju *ground truth* labela klastera.
  - Računa se **v mera** (eng. *v measure score*), koja se zasniva na intuitivnim metrikama zasnovanim na uslovnoj analizi entropije:
    - **homogenost** (eng. *homogenety*) - svaki klaster sadrži članove samo jedne grupe/klastera
    - **potpunost** (eng. *completeness*) - svi članovi iste grupe/klastera su dodeljeni istom klasteru.
  - **v mera** predstavlja harmonijsku sredinu homogenosti i potpunosti:
    - *sklearn.metrics.v_measure_score(labels_true, labels_pred)*

# Zadatak 5

- Saveti za rešavanje zadatka:
  - Podsetiti se gradiva sa predavanja
  - Uraditi eksplorativnu analizu podataka
  - Isprobati više tehnika za rad sa nedostajućim vrednostima
  - Primeniti klasterovanje i analizirati klastere.