



# Mašinsko učenje 2020

Zadatak 6



# Sadržaj

- Zadatak 5 - Rekapitulacija
- Zadatak 6



## **Zadatak 5 - Rekapitulacija**



## Zadatak 5 - Rekapitulacija

- Procenat uspešnosti: **69.7%** (23/33).
- Sva rešenja sa **v measure  $\geq 0.6$**  se smatraju odličnim.
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **18%**.

- Najbolji rezultati po terminima:

Termin	Tim	v measure
Ponedeljak	<b>mjau_mace</b>	<b>0.749</b>
Utorak 1	<b>tim8_20</b>	<b>0.673</b>
Utorak 2	<b>demo</b>	<b>0.683</b>
Četvrtak	<b>sljub_duo</b>	<b>0.653</b>
Petak	<b>masinski_ucenjaci</b>	<b>0.70</b>



## Zadatak 5 - Rekapitulacija

- Dobre stvari (na nivou generacije):
  - Rad sa nedostajućim vrednostima;
  - Određivanje hiperparametara modela;
  - Prpratni izveštaji.



# Zadatak 6



## Zadatak 6

- PCA + Klasifikacija:
  - Na osnovu dostupnih informacija o zaposlenima na istočnoj obali SAD, izvršiti predikciju njihove rase (**race**):
    - 1. White
    - 2. Black
    - 3. Asian
    - 4. Other



# Zadatak 6

- PCA + Klasifikacija:
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije mikro f1 mera (eng. *micro f1 score*)  $> 0.79$ .
  - **Obavezna upotreba PCA!**
  - Dozvoljena upotreba svih rađenih klasifikatora/ansambla klasifikatora.
  - Rok: **07.06.2020. u 12:59h.**
  - Trening skup podataka sadrži nedostajuće vrednosti (prazne ćelije).
  - Instalirane biblioteke za Zadatak 6 (verzije date u Uputstvu):
    - NumPy
    - SciPy
    - Pandas
    - scikit-learn.





# Zadatak 6

- Atributi:
  - **year** - godina kada su prikupljane informacije
  - **age** - broj godina (starost) zaposlenih
  - **maritl** - bračni status zaposlenih:
    - 1. **Never Married** - nikad venčani
    - 2. **Married** - venčani
    - 3. **Widowed** - udovice/udovci
    - 4. **Divorced** - razvedeni
    - 5. **Separated** - rastavljeni



# Zadatak 6

- Atributi:
  - **education** - nivo obrazovanja:
    - 1. < **HS Grad** - nezavršena srednja škola
    - 2. **HS Grad** - završena srednja škola
    - 3. **Some College** - nezavršen fakultet
    - 4. **College Grad** - završen fakultet
    - 5. **Advanced Degree** - MSc, PhD
  - **jobclass** - tip posla:
    - 1. **Industrial** - industrijski
    - 2. **Information** - informacioni



## Zadatak 6

- Atributi:
  - **health** - zdravstveno stanje
    - 1. **<= Good** - dobro ili slabije
    - 2. **>= Very Good** - veoma dobro ili odlično
  - **health\_ins** - da li zaposleni poseduje zdravstveno osiguranje:
    - 1. **Yes**
    - 2. **No**
  - **wage** - godišnja plata (u hiljadama \$).



## Zadatak 6

- Trening skup podataka sadrži nedostajuće vrednosti (u pitanju su prazne ćelije).
  - Testni skup podataka **ne** sadrži nedostajuće vrednosti.
- 
- Prilikom izrade rešenja, **obavezna** je upotreba PCA.
  - PCA možete implementirati samostalno, a možete iskoristiti i implementacije iz scikit-learn biblioteke:
    - [PCA](#)
    - [KernelPCA](#)