Рачунарство високих перформанси у информационом инжењерингу

Опште информације за практични део

Потребно је анализирати један скуп података спроводећи активности организоване према следећим целинама:

• припрема података [3п]

- учитавање података
- уређивање учитаних података
- испитивање присуства недостајућих вредности

• прелиминарна анализа података [3п]

- израчунавање вредности дескриптивних статистика по појединачним обележјима
- визуализовање расподеле по појединачним обележјима
- испитивање односа између обележја

класификација [15п]

- формирање класификационих модела за исто циљно обележје применом три различита метода класификације и оцењивање перформанси у класификацији
 - за сваки метод класификације формирање класификационих модела за различите вредности параметара према три сценарија
 - израчунавање вредности различитих бројчаних показатеља перформанси
 - одређивање перформанси применом унакрсне валидације
 - испитивање односа између вредности параметара и перформанси у класификацији
 - бирање решења за класификацију по сваком од метода и на нивоу свих метода

кластеризација [6п]

- р формирање кластера применом једног метода кластеризације
 - формирање кластера̂ за различите вредности параметара према два сценарија
 - испитивање структуре добијених кластера
 - визуализовање односа између вредности обележја и припадности кластеру

извештавање [3п]

- формирање аналитичког извештаја
 - специфицирање коришћених података и технологија
 - приказивање свих радњи над подацима кроз програмски ко̂д
 - приказивање свих остварених резултата и тумачење тих резултата

Очекивана је употреба софтверских технологија намењених раду на великим количинама података кроз језик *R* и софтверске технологије за извештавање кроз језик *Markdown*. Скуп података је барем реда величине *GB*, довољно разноврстан у погледу природе заступљених обележја за смислену примену у класификацији и кластеризацији, слободно доступан без процедура за остварење приступа и слободно расположив за употребу.