Оцена знања за предмет „Увод у науку о подацима“

Начин полагања испита:

- **Предиспитне обавезе: 70 поена**
  - **Колоквијум** – 24 поена
  - **Семинарски рад** – 46 поена
- **Завршни испит: 30 поена**

Услов за израду семинарског рада је да студент освоји минимално 12 поена на колоквијуму.

Семинарски рад се ради у тимовима од 2 или 3 члана. Тимови од 3 члана додатно постављају моделе у продукцију.
Одбрана семинарских радова укључује и питања из праксе и питања из теорије.

Услов за излазак на завршни део испита је освојених 36 поена на предиспитним обавезама.

Семинарски рад

Поред колоквијума, предиспитне обавезе из предмета *Увод у науку о подацима* стичу се анализом једне базе података доступне на сајту kaggle.com/datasets. Можете да користите и друге ресурсе који дају податке.

За оне најхрабрије, тема се може изабрати и са листе завршних пројеката курса Машинског учења Станфорд универзитета. Свака тема је покривена радом који описује процес који треба опонашати и укључује одговарајући скуп података и пратеће кодове. Ови семинарски радови се високо котирају!

Кораци

- Студенти бирају базу података према сопственим афинитетима. Када се одлучите за скуп података, најпре се професору шаље емаил како би изабрана база података била **одобрена**.

  Садржај емаила: Маилом се шаље мањи документ са крађим описом скупа података, колико редова и колона има, да ли има недостајуће вредности, коју колону желите да предвиђате – њену расподелу. Додатно можете додати и друге ствари које сматрате релевантним.

  Након тога, на Teams-у постављате линк на базу података којим обавештавате остале колеге у вези Вашег избора. Различити тимови треба да анализирају различите базе података.

- Семинарски рад треба да садржи:

  1. Базу података припремљену за R/Python-програмски језик у виду једног или више оквира података у којима је решен проблем непостојећих или нетипичних вредности променљивих. Анализа база података које немају недостајуће или нетипичне вредности утиче на коначну оцену семинарског рада.

  2. Извештај у **.pdf** или **.docx** формату који представља оргиналан резултат анализе одабраних података. Податке анализирати R/Python -програмским апаратима, а опсервације представити графички и ако је могуће потврдити одговарајућим статистичким тестовима са подразумеваним прагом значајности 0,05 (детаље везане за статистичке тестове можете да видите на линку: https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf). По потреби оквиру података додати нове променљиве (*feature engineering*).

     Први део извештаја треба да садржи кратак опис базе која се анализира као и структуру података који су на располагању.

     Сваки корак у анализи треба да садржи:

     2.1 мотивацију или образложење за апарат који се користи;

     2.2 код у R/Python-у којим се апарат извршава;

     2.3 резултат извршења R/Python -команде;

     2.4 тумачење добијених резултата.

     На крају извештаја написати закључак о добијеним опсервацијама.

3. **R-markdown** или **Python нотебук** који хронолошки прати све извршене команде у извештају.

4. Након добијених резултата, испробати неколико алгоритама машинског учења по жељи и протумачити резултате. Податке поделити на скупове за тренирање и тестирање.

- Скуп података и „чист“ код морају да буду на вашим **гитхаб/гитлаб налозима** (ово ће вам треба прилико писања биографије и аплицирања за посао или праксу).

ДОДАТНЕ НАПОМЕНЕ: Метрике треба применити искључиво на тестном скупу, а пожељно је и коришћење крос-валидације. Уколико резултати теста нису задовољавајући, онда треба приступити новим стратегијама припреме података у циљу побољшања модела. Циљ је не само показати како се препроцесирају подаци, већ и направити робустан модел који је у стању да упореби знање из тих података.

Неке стратегије израде семинарског рада могу бити, на пример, другачији приступ за решавање проблема недостајућих података; додавање нових карактеристике (*features*), неке карактеристике се потенцијално могу избацити (feature selection), неке трансформисати ако има потребе итд; проблем небалансираних података; направити претрагу параметара за алгоритам...

Студенти треба да предају семинарски рад најмање 5 дана пре полагања испита у року који одаберу. Испит подразумева одбрану семинарског рада као и проверу знања из предмета.

За евентуалне консултације јавити се мејлом на branko.arsic@pmf.kg.ac.rs.