

# Inteligentná analýza údajov, letný semester 2017/18

## Hodnotenie a podmienky absolvovania predmetu

### Priebežné hodnotenie

- projekt: **max. 40 bodov**
  - práca na projekte a úlohách v rámci cvičení: 7 bodov
  - 1. priebežná správa - prieskumná analýza (v 6. týždni): 10 bodov
  - 2. priebežná správa - predspracovanie údajov (v 9. týždni): 15 bodov
  - výsledná správa (v 12. týždni): 8 bodov
- Projekt bude riešený vo dvojiciach.

### Záverečné hodnotenie

- body získané počas semestra: max. 40 bodov
- skúška: max. 60 bodov

### Podmienky na získania zápočtu

1. vypracovanie všetkých častí projektu v akceptovateľnej kvalite, jeho odovzdanie a prezentovanie podľa harmonogramu
2. aktívna účasť na cvičeniach
3. získanie aspoň 25 bodov počas semestra

### Neskoré odovzdanie

V prípade nedodržania termínu na odovzdanie jednotlivých fáz zadania do systému AIS bude možné danú fázu odovzdať do siedmich dní s 50% penalizáciou. Neskoršie odovzdanie nebude možné. Neodovzdanie niektorej z fáz projektu bude mať za následok neudelenie zápočtu.

### Podmienky absolvovania predmetu

Pre absolvovanie predmetu platia všeobecné podmienky absolvovania a hodnotenia predmetov. Za samozrejmosť sa považuje dodržanie pravidiel akademickkej korektnosti. Akékoľvek vydávanie cudzej práce za vlastnú je neakceptovateľné, a to v akomkoľvek rozsahu. Ak sa ho dopustíte, automaticky to znamená neúspešné ukončenie predmetu (FX) a podnet na disciplinárnu komisiu.

# Projekt

Cieľom projektu je osvojiť si základné koncepty a techniky analýzy dát, pochopiť, ako fungujú a získať intuíciu pre ich vhodnú aplikáciu za účelom objavovania znalostí v dátach. Mali by ste tiež získať predstavu, aké otázky vieme pomocou analýzy dát zodpovedať a byť schopní aplikovať a vyhodnotiť základné prístupy strojového učenia.

**Projekt sa vypracúva vo dvojiciach.** Pri riešení sa očakáva využitie jazyka Python a dostupných knižníc na analýzu dáta (pandas, matplotlib a pod.). V každej fáze sa odovzdáva vykonateľný Jupyter Notebook, ktorý by mal zachytávať a vhodne dokumentovať všetky vykonané transformácie nad dátami. Odovzdaný notebook musí obsahovať nielen kód, ale aj jeho výsledky (vypočítané hodnoty, výpisy, vizualizácie a pod.) spolu s komentárom intepretujúcim získané výsledky z toho plynúce rozhodnutia pre ďalšie kroky dátovej analýzy. Schopnosť dobre odkomunikovať a vybrať relevantné výsledky analýzy bude predstavovať významnú zložku hodnotenia.

## Dáta

Každá dvojica bude pracovať s im náhodne pridelenou dátovou sadou. Dáta predstavujú záznamy o pacientoch s chorobou štítnej žľazy. Vašou úlohou je vedieť predikovať hodnotu **Y**. Budete sa musieť pritom vysporiadať s viacerými problémami, ktoré sa v dátach nachádzajú (formáty dát, chýbajúce, nezmyselné alebo vychýlené hodnoty a pod.).

## Prieskumná analýza (max. 10b)

Prieskumná analýza je kľúčovou časťou analýzy dát. Bez nej nie sme schopní dáta spracúvať, pretože nevieme, čo sa v nich nachádza. Využíva sa pritom predovšetkým deskriptívna štatistika a rôzne podporné vizualizácie.

V tejto fáze sa od vás očakáva:

- **Opis dát spolu s ich charakteristikami** (počet záznamov, počet atribútov, ich typy, distribúcie, základné deskriptívne štatistiky a pod.).
- **Identifikácia problémov v dátach spolu s predpokladaným scenárom riešenia v ďalšej fáze**, t. j., čo budete musieť v rámci predspracovania vyriešiť (aj s naznačením možností, ako tieto problémy plánujete v ďalšej fáze riešiť). Medzi problémy, na ktoré môžete v rámci analýzy naraziť, patria napr.:
  - nevhodná štruktúra dát (dáta nie sú v tabuľkovej podobe alebo jedna entita je opísaná viacerými riadkami tabuľky)
  - nejednotné formáty dát
  - chýbajúce hodnoty
  - vychýlené (odľahlé) hodnoty
  - a ďalšie, t. j. v dátach sa môžu nachádzať aj iné, tu nevymenované problémy, ktoré tiež treba identifikovať a vo vašej analýze adresovať.

V odovzdanej správe (Jupyter Notebooku) by ste tak mali vedieť zodpovedať na otázky:

- Majú dáta vhodný formát pre ďalšie spracovanie? Ak nie, aké problémy sa v nich vyskytujú?
- Sú v dátach chýbajúce hodnoty? Ako sú reprezentované? Ako plánujete riešiť problém chýbajúcich hodnôt pre jednotlivé atribúty, resp. pozorovania? (Pre rôzne atribúty môže byť vhodné použiť rôzne stratégie.)
- Nadobúdajú niektoré atribúty nezmyselné (nekonzistentné) či inak výrazne odchylené hodnoty? Ktoré?
- Ako plánujete v ďalšej fáze tieto identifikované problémy adresovať / riešiť?

Správa sa odovzdáva v 6. týždni semestra na cvičení (dvojica svojmu cvičiacemu odprezentuje vykonanú prieskumnú analýzu v Jupyter Notebooku). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do **nedele 25.3.2018 do 23:59**.

## Predspracovanie (max. 15b)

Na základe identifikovaných problémov v dátach a návrhu ich riešenia v predchádzajúcej fáze treba zrealizovať predspracovanie. Výsledkom by mala byť upravená dátová sada (vo formáte csv) vo vhodnom tvare pre zvolený algoritmus strojového učenia (v našom prípade *rozhodovacie stromy*). Zároveň, keďže predspracovaním sa mohol zmeniť tvar a charakteristiky dát (počet atribútov, distribúcie hodnôt a pod.), treba znovu zrealizovať podstatné časti prieskumnej analýzy, pričom v tejto fáze je potrebné zamerať sa aj na vzťahy medzi jednotlivými atribútmi.

V tejto fáze sa od vás očakáva:

- **Realizácia krokov predspracovania dát a ich zdokumentovanie.**
  - Pri riešení chýbajúcich hodnôt vyskúšajte rôzne stratégie (očakáva sa vyskúšanie a porovnanie *minimálne dvoch* stratégií, pričom aspoň jedna z nich musí byť zvolená z posledných troch menovaných):
    - nahradenie chýbajúcej hodnoty mediánom
    - nahradenie chýbajúcej hodnoty priemerom
    - nahradenie chýbajúcej hodnoty pomerom ku korelovanému atribútu
    - nahradenie chýbajúcej hodnoty priemerom segmentu
    - nahradenie chýbajúcej hodnoty pomocou lineárnej regresie
    - nahradenie chýbajúcej hodnoty pomocou algoritmu k-najbližších susedov
  - Podobne postupujte aj pri riešení vychýlených (odľahlých) hodnôt, pričom porovnajte odstránenie vychýlených pozorovaní s aspoň jednou zvolenou stratégiou ich nahradenia, resp. transformácie:
    - odstránenie vychýlených (odľahlých) pozorovaní
    - nahradenie vychýlenej hodnoty hraničnými hodnotami rozdelenia (5 percentilom, resp. 95 percentilom)
    - transformácia atribútu s vychýlenými hodnotami pomocou zvolenej funkcie (logaritmus, odmocnina a pod.)
- **Opätovná realizácia podstatných častí prieskumnej analýzy.** V rámci nej by ste mali vedieť zodpovedať na otázky:

- Ako sa zmenili distribúcie hodnôt po realizácii krokov predspracovania?
- Sú niektoré atribúty medzi sebou závislé?
- Od ktorých (jednotlivých) atribútov závisí predikovaná premenná?
- **Manuálne vytvorenie a vyhodnotenie rozhodovacích pravidiel pre klasifikáciu.** Vyskúšajte jednoduché pravidlá zahŕňajúce jeden atribút, ale aj komplikovanejšie zahŕňajúce viacero atribútov (ich kombinácie). V tejto fáze sa neočakáva použitie žiadneho algoritmu strojového učenia, pravidlá by mali byť vytvorené manuálne na základe pozorovaných závislostí v dátach. Pravidlá (manuálne vytvorené klasifikátory) vyhodnoťte pomocou metrík *správnosť* (angl. *accuracy*), *presnosť* (angl. *precision*) a *úplnosť* (angl. *recall*).

Správa sa odovzdáva v 9. týždni semestra na cvičení (dvojica svojmu cvičiacemu odprezentuje vykonané predspracovanie v Jupyter Notebooku). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do **nedele 15.4.2018 do 23:59**.

## Strojové učenie (max. 8b)

Pri dátovej analýze nemusí byť našim cieľom získať len znalosti obsiahnuté v aktuálnych dátach, ale aj natrénovať model, ktorý bude schopný robiť rozumné predikcie pre nové pozorovania (v našom prípade pre nových pacientov). Na to sa využívajú techniky strojového učenia. V tomto projekte sa zameriame na *rozhodovacie stromy* vzhľadom na ich jednoduchú interpretovateľnosť.

V poslednej fáze sa od vás očakáva:

- **Natrénovanie klasifikátora s využitím rozhodovacích stromov.** Využite algoritmus dostupný v knižnici *scikit-learn* (CART). Preskúmajte hyperparametre tohto algoritmu a vyskúšajte ich rôzne nastavenie tak, aby ste minimalizovali preučenie. Vysvetlite, čo jednotlivé hyperparametre robia. Pri nastavovaní hyperparametrov algoritmu využite 10-násobnú krížovú validáciu na trénovacej množine.
- **Vyhodnotenie natrénovaných klasifikátorov.** Vizualizujte natrénované pravidlá. Porovnajte algoritmy navzájom, ako aj s vašimi manuálne vytvorenými pravidlami z predchádzajúcej fázy. Vyhodnoťte ich pomocou metrík *správnosť* (angl. *accuracy*), *presnosť* (angl. *precision*) a *úplnosť* (angl. *recall*).
- **Vyhodnotenie vplyvu zvolenej stratégie riešenia chýbajúcich hodnôt na *správnosť* klasifikácie.** Zistite, či použitie zvolených stratégií riešenia chýbajúcich hodnôt vplyva na *správnosť* (angl. *accuracy*) klasifikácie. Ktorá stratégia sa ukázala ako vhodnejšia pre daný problém?

Správa sa odovzdáva v 12., resp. 13. týždni semestra na cvičení (dvojica svojmu cvičiacemu odprezentuje vykonanú prieskumnú analýzu v Jupyter Notebooku). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do **nedele 13.5.2018 do 23:59**.