Úvod

Cieľom projektu je osvojiť si **prehľad fungovania v dátovej vede**, základné koncepty a techniky analýzy dát, pochopia, ako fungujú a získajú intuíciu pre ich vhodnú aplikáciu za účelom objavovania znalostí v dátach. Taktiež získajú predstavu, aké otázky vieme pomocou analýzy dát zodpovedať a aplikovať **základné prístupy strojového učenia**. Dôraz je kladený na analýzu a predspracovanie dát, použitie metód strojového učenia, spôsoby ich vyhodnotenia a porovnania.

Projekt sa vypracúva **v dvojiciach**. Pri riešení sa používa programovací jazyk **Python** a dostupné knižnice pre dátovú vedu ako **pandas, numpy, scipy, statsmodels, scikit-learn,** atd.. V každej fáze sa odovzdáva vykonateľný **Jupyter Notebook** do AISu, ktorý obsahuje všetky vykonané transformácie nad dátami s vhodnou dokumentáciou. Odovzdaný notebook musí obsahovať nielen kód, ale aj jeho výsledky (vypočítané hodnoty, výpisy, vizualizácie a pod.) spolu s komentárom k získaným výsledkom a z toho plynúce rozhodnutia pre ďalšie kroky dátového procesu. Schopnosť dobre komunikovať a prezentovať relevantné výsledky predstavuje významnú zložku hodnotenia.

Pri každej fáze v odovzdanom notebooku uveďte **percentuálny podiel práce** členov dvojice.

Dáta (Data)

https://drive.google.com/drive/folders/1joPJW_nVLUKXKeFDWneFgEfvgpbRe15z?usp=sharing
(každá dvojica má jeden dataset pod číslom, ktoré ste si vybrali na cvičení)

V dnešnej dobe sa nakupovanie cez internet stalo bežnou súčasťou života veľkej časti populácie. Používatelia si môžu vybrať z veľkého množstva produktov či vedia efektívne porovnávať kvalitu a cenu. Avšak nie každý používateľ, ktorý navštívi e-obchod v ňom aj nakúpi, a preto rozpoznanie nákupného správania zákazníka zohráva dôležitú úlohu. Ak systém včas identifikuje zákazníka, ktorý sa rozhodol napokon nenakúpiť, vie mu vygenerovať napr. zľavu, alebo ponúknuť iné relevantné produkty. Je žiaduce modelovať správanie používateľa na základe jeho interakcií s e-obchodom. V záznamoch je závislá premenná s menom **“*ack*”** indikujúca nákup zákazníka počas jedného sedenia (session) v e-obchode. Táto premenná je zaznamenaná ako poďakovanie e-obchodu (e-shop) zákazníkovi (user) po zaplatení za továr (product).

Zadanie (The quest)

Každá dvojica bude pracovať s pridelenou dátovou sadou od 3. týždňa. **Vašou úlohou** je predikovať závislé hodnoty premennej “***ack***” (predikovaná premenna) pomocou metód strojového učenia. Budete sa musieť pritom vysporiadať s viacerými problémami, ktoré sa v dátach nachádzajú ako formáty dát, chýbajúce, vychýlené hodnoty a mnohé ďalšie. Očakavaným **výstupom** projektu je:

1. **najlepší model** strojového učenia;
2. **data pipeline** pre jeho vybudovanie na základe vstupných dát.

Popis produktov: premenné (variable, feature), ktoré sa môžu vyskytnúť v datasetoch

product_ean
store_name
code
location

Popis používateľa: premenné (variable, feature), ktoré sa môžu vyskytnúť v datasetoch

user_id
username
name
address
current_location
residence
birthdate
registration
race
sex
mail
job

Popis sedenia (session): premenné (variable, feature), ktoré sa môžu vyskytnúť v datasetoch

session_id
session_start
session_duration
total_load_time
screen_width
screen_height
browser_name

Interakcie počas sedenia

page_activity_duration	trvanie aktivity používateľa na stránke
wild_mouse_duration	trvanie rýchleho pohybu myši,
mouse_move_total_rel_distance	normalizovaná na výšku a šírku obrazovky
scroll_move_total_rel_distance	normalizovaná na výšku a šírku obrazovky

(pct_) pomer počtu interakcií daného typu k celkovému počtu nasobený bázovou hodnotou

pct_scroll_move_duration	normonovaná hodnota na celkový čas rolovania
pct_mouse_move	záznamy o pohybe myši
pct_scroll_move	záznamy o rolovaní
pct_wild_mouse	záznamy o rýchlom pohybe myši
pct_click	záznamy o klikaní používateľa
pct_double_click	záznamy o dvojitom kliknutí používateľa
pct_rage_click	záznamy o zúrivom kliknutí na stránke
pct_input	záznamy o zadávaní vstupov používateľom
pct_scrandom	záznamy o rýchlom presúvaní na stránke
pct_click_product_info	záznamy o klikaní na informácie produktu

Popis sedenia, zoznam interakcií, (pct_) sú redukované z väčšieho množstva monitorovaných popisov a interakcií v e-shope. Originálne dáta (logy) majú veľkosť GB v formáte JSON.

Fáza 1 - Prieskumná analýza (v 6. týždni): 15% = 15 bodov

Základný opis dát spolu s ich charakteristikami (5b)

EDA s vizualizáciou

- (1b) Analýza štruktúr dát:
  - súbory: štruktúry a vzťahy, počet, typy, ...
  - Záznamy: štruktúry, počet záznamov, počet atribútov, typy, ...
- (1b) Analýza jednotlivých atribútov: pre zvolené významné atribúty (min 10) analyzujte ich distribúcie a základné deskriptívne štatistiky.
- (1b) Párová analýza dát: Identifikujte vzťahy a závislostí medzi dvojicami atribútov, napr. korelácie.
- (1b) Párová analýza dát: Identifikujte závislosti medzi predikovanou premennou a ostatnými premennými (potenciálnymi prediktormi).
- (1b) Dokumentujte Vaše prvotné zamyslenie k riešeniu zadania projektu, napr. sú niektoré atribúty medzi sebou závislé? od ktorých atribútov závisí predikovaná premenná? či sa dá kombinovať záznamy z viacerých súborov? či je potrebné ich kombinovať?

Identifikácia problémov v dátach s prvotným riešením (5b)

- (3b) Identifikujte problémy v dátach napr.: nevhodná štruktúra dát, duplicitné záznamy (riadky, stlpce), nejednotné formáty, chýbajúce hodnoty, vychýlené hodnoty. V dátach sa môžu nachádzať aj iné, tu nevymenované problémy.
- (2b) Navrhnuté riešenie problémov s dátami prvotne realizujte na dátach. Problémy s dátami môžete riešiť iteratívne v každej fáze aj vo všetkých fázach podľa Vašej potreby.

Formulácia a štatistické overenie hypotéz o dátach (5b)

- (5b) Sformulujte **dve hypotézy** o dátach v kontexte zadanej predikčnej úlohy. Formulované hypotézy overte vhodne zvolenými štatistickými testami.
  Príklad formulovania hypotézy:
  trvanie rýchleho pohybu myši (wild_mouse_duration) má v priemere vyššiu hodnotu pri nákupe (ack=1) ako bez nákupu (ack=0) počas sedenia.

V odovzdanej správe (Jupyter notebook) by ste tak mali odpovedať na otázky:

- Majú dáta vhodný formát pre ďalšie spracovanie? Ak nie, aké problémy sa v nich vyskytujú?
- Sú v dátach chýbajúce hodnoty? Ako plánujete riešiť tento problém?
- Nadobúdajú niektoré atribúty nekonzistentné alebo výrazne odchýlené hodnoty?
- Ako plánujete/riešíte tieto identifikované problémy?

Správa sa odovzdáva v 6. týždni semestra. Dvojica svojmu cvičiacemu odprezentuje vykonanú fázu v Jupyter Notebooku podľa potreby na cvičení. V notebooku uveďte **percentuálny podiel práce** členov dvojice. Následne správu elektronicky odovzdá **jeden člen z dvojice** do systému **AIS** do nedele **29.10.2023 23:59**.

Fáza 2 - Predspracovanie údajov (v 9. týždni): 20 bodov

V tejto fáze sa od Vás očakáva že realizujte **predspracovanie údajov** pre strojové učenie. Výsledkom bude dátová sada (csv alebo tsv), kde jedno pozorovanie je opísané jedným riadkom.

- **scikit-learn** vie len numerické dáta, takže treba niečo spraviť s nenumerickými dátami.
- Replikovateľnosť predspracovania na trénovacej a testovacej množine dát, aby ste mohli zopakovať predspracovanie viackrát podľa Vašej potreby (iteratívne).

Keď sa predspracovaním mohol zmeniť tvar a charakteristiky dát, je treba realizovať EDA opakovane podľa Vašej potreby. Bodovať znovu (EDA) nebudeme, zmeny ale dokumentujte. Problém s dátami môžete riešiť iteratívne v každej fáze aj vo všetkých fázach podľa potreby.

Integrácia a čistenie dát (5b)

Transformujte dáta na vhodný formát pre strojové učenie t.j. jedno pozorovanie musí byť opísané jedným riadkom a každý atribút musí byť v numerickom formáte.

- (2b) Chýbajúce hodnotu (missing values): vyskúšajte min. 2 techniky ako napr.
  - odstránenie pozorovaní s chýbajúcimi údajmi
  - nahradenie chýbajúcej hodnoty mediánom, priemerom, pomerom (ku korelovanému atribútu), alebo pomocou lineárnej regresie resp. kNN
- (3b) Podobne pri riešení vychýlených hodnôt (outlier detection), min. 2 techniky napr.
  - odstránenie vychýlených (odľahlých) pozorovaní
  - nahradenie vychýlenej hodnoty hraničnými hodnotami rozdelenia (napr. 5%, 95%)

Realizácia predspracovania dát (5b).

- (1b) Transformované dáta pre ML si rozdeľte na trénovaciu a testovaciu množinu podľa vami preddefinovaného pomeru. Ďalej pracujte len **s trénovacím datasetom**.
- (3b) Transformujte atribúty dát pre strojové učenie podľa dostupných techník (minimálne 3 techniky) ako scaling, transformers a ďalšie.
- (1b) Zdôvodnite Vaše voľby/rozhodnutie pre realizáciu (t.j. zdokumentovanie)

Výber atribútov pre strojové učenie (5b)

- (2b) Zistite, ktoré atribúty (features) vo vašich dátach pre strojové učenie sú informatívne k predikovanej premennej (minimálne 2 techniky s porovnaním medzi sebou).
- (1b) Zoradíte zistené atribúty v poradí podľa dôležitosti.
- (2b) Zdôvodnite Vaše voľby/rozhodnutie pre realizáciu (t.j. zdokumentovanie)

Replikovateľnosť predspracovania (5b)

- (3b) Upravte váš kód realizujúci predspracovanie trénovacej množiny tak, aby ho bolo možné bez ďalších úprav znovu použiť **na predspracovanie testovacej množiny** (napr. pomocou funkcie/í) v kontexte strojového učenia.
- (2b) Očakáva sa aj využitie možnosti **sklearn.pipeline**

Správa sa odovzdáva v 9. týždni semestra. Dvojica svojmu cvičiacemu odprezentuje vykonanú fázu v notebooku podľa potreby na cvičení. Uveďte percentuálny podiel práce členov dvojice. Následne správu elektronicky odovzdá **jeden člen z dvojice** do systému **AIS** do nedele **19.11.2023 23:59**.

Fáza 3 - Strojové učenie (v 12. týždni): 20 bodov

Pri dátovej analýze nemusí byť naším cieľom získať len znalosti obsiahnuté v aktuálnych dátach, ale aj natrénovať model, ktorý bude schopný robiť rozumné **predikcie** pre nové pozorovania pomocou techniky **strojového učenia**.

Jednoduchý klasifikátor na základe závislosti v dátach (5b)

- (3b) Naimplementujte OneR algorithm (iné mená: OneRule or 1R), ktorý je jednoduchý klasifikátor - rozhodnutie na základe jedného atribútu. Algoritmus má byť realizovaný na základe závislostí v dátach.
- (1b) Realizujte rozhodnutie na základe viac atribútov pomocou Vášho OneR.
- (1b) Vyhodnoťte klasifikátor (rozhodnutie na základe jedného atribútu, rozhodnutie na základe viac atribútov) pomocou metrík accuracy, precision a recall.

Trénovanie a vyhodnotenie klasifikátorov strojového učenia (5b)

- (1b) Na trénovanie využite **minimálne jeden stromový algoritmus v scikit-learn**.
- (1b) Porovnajte **minimálne** s jedným iným algoritmom **v scikit-learn**.
- (1b) Porovnajte s Vaším OneR z prvého kroku.
- (1b) Vizualizujte natrénované pravidlá **minimálne** pre jeden Vami vybraný algoritmus
- (1b) Vyhodnoťte natrénované modely pomocou metrík accuracy, precision a recall

Optimalizácia alias hyperparameter tuning (5b)

- (1b) Preskúmajte aspoň jeden Vami vybraný klasifikačný algoritmus v druhom kroku detailnejšie. Vysvetlite jeho hyperparametre a prečo ste si ho vybrali.
- (2b) Vyskúšajte rôzne nastavenie hyperparametrov (tuning) a kombinácie modelov (ensemble) pre ten zvolený algoritmus tak, aby ste **minimalizovali overfitting** (preučenie) a **optimalizovali výkonnosti**.
- (2b) Pri nastavovaní modelu využite **krížovú validáciu** (cross validation) na trénovacej množine (stabilita modelu).

Vyhodnotenie vplyvu zvolenej stratégie riešenia na klasifikáciu (5b)

Vyhodnoťte Vami zvolené stratégie riešenia projektu z hľadiska classification accuracy:

- (1b) Stratégie riešenia chýbajúcich hodnôt a outlierov;
- (1b) Či dátová transformácia (scaling, transformer, ...) zlepší accuracy klasifikácie;
- (1b) Výber atribútov a výber algoritmov strojového učenia, či majú vplyv na výkonnosť (presnosť a run-time)
- (1b) Hyperparameter tuning resp. ensemble learning.
- (1b) Ktoré spôsoby z hore-uvedených bodov sa ukázali ako učinné pre Váš dataset? Ktorý model je Váš **najlepší model** pre nasadenie (deployment)? Aký je **data pipeline** pre jeho vybudovanie na základe Vášho datasetu v produkcii?

Všetky hodnotenia podložte dôkazmi. Najlepší model má byť stabilný, bez overfitu a bez underfitu. Jeho data pipeline má byť dodaný s metadátami, ak sú potrebné a vyrobené v developmente.

Správa sa odovzdáva v poslednom týždni semestra. Dvojica svojmu cvičiacemu odprezentuje vykonanú fázu v Jupyter Notebooku podľa potreby na cvičení. V notebooku uveďte percentuálny podiel práce členov dvojice. Následne správu elektronicky odovzdá **jeden člen z dvojice** do systému **AIS** do **12.12.2022 23:59**.