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** v akceptovateľnej kvalite. Pri riešení sa používa programovací jazyk **Python** a dostupné knižnice pre dátovú vedu ako **pandas, numpy, scipy, statsmodels, scikit-learn,** atd.. V každej fáze a aktivite 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.

Data

https://drive.google.com/drive/folders/1vLIh5f3ix4KGQm0qX1cj2uGUVg0G7r5T?usp=share_link
(každá dvojica má jeden dataset pod číslom, ktoré máte na cvičení)

Mobilné zariadenia sú dnes neoddeliteľnou súčasťou interakcie človeka s počítačom (HCI Human-Computer Interaction), pretože poskytujú používateľom rýchly a pohodlný prístup k informáciám a aplikáciám. Tento druh interakcie je intuitívny a efektívny, no zároveň otvára dvere pre hrozby, ako je malware. Malware, čiže škodlivý softvér, dokáže infiltrovať mobilné zariadenia prostredníctvom infikovaných aplikácií alebo škodlivých webových stránok, pričom môže kradnúť citlivé údaje alebo získať kontrolu nad zariadením. Aby boli mobilné zariadenia bezpečné, je kľúčové pre inteligentné antimalvérové softvéry rychlé a presné detegovanie a následne včasné varovanie užívateľom v čo najkratšom čase. Jadro takýchto softvérov je vybudovaný na základe poskytnutých dátach tzv. záznamy (angl. logy) a detegovanie sa robí pomocou strojového učenia.

V záznamoch (dataset pre Vás) je závislá premenná s menom **“*mwra*”** indikujúca *malware-related-activity* v jednom časovom intervale. Dataset je zalohovaný pomocou Rapid7 agenta (https://www.rapid7.com) nasadeného na androidových mobilných zariadeniach (https://developer.android.com). Dataset je čiastočne predspracovaný pre IAU projektový učel.

Zadanie (The QUEST)

Každá dvojica bude pracovať s pridelenou dátovou sadou od 2. týždňa. **Vašou úlohou** je predikovať závislé hodnoty premennej “***mwra***” (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.

Fáza 1 - Prieskumná analýza: 15% = 15 bodov

1.1 Základný opis dát spolu s ich charakteristikami (5b)

EDA s vizualizáciou

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

1.2 Identifikácia problémov, integrácia a čistenie dát (5b)

- (A-2b) Identifikujte aj prvotne riešte 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.
- (B-2b) Chýbajúce hodnoty (missing values): vyskúšajte riešiť problém min. 2 technikami
    - odstránenie pozorovaní s chýbajúcimi údajmi
    - nahradenie chýbajúcej hodnoty napr. mediánom, priemerom, pomerom, interpoláciou, alebo kNN
- (C-1b) Vychýlené hodnoty (outlier detection), vyskúšajte riešiť problém min. 2 technikami
    - odstránenie vychýlených alebo odľahlých pozorovaní
    - nahradenie vychýlenej hodnoty hraničnými hodnotami rozdelenia (napr. 5%, 95%)

1.3 Formulácia a štatistické overenie hypotéz o dátach (5b)

- (A-4b) Sformulujte **dve hypotézy** o dátach v kontexte zadanej predikčnej úlohy. Formulované hypotézy overte vhodne zvolenými štatistickými testami.
  Príklad formulovania:
  > *android.defcontainer má v priemere vyššiu váhu v stave malware-related-activity ako v normálnom stave*
- (B-1b) Overte či Vaše štatistické testy majú dostatok podpory z dát, teda či majú dostatočne silnú štatistickú silu.

V odovzdanej správe (Jupyter notebook) by ste tak mali odpovedať na otázky:
Majú dáta vhodný formát pre ďalšie spracovanie? Aké problémy sa v nich vyskytujú? Nadobúdajú niektoré atribúty nekonzistentné hodnoty? Ako riešíte tieto Vami identifikované problémy?

Správa sa odovzdáva v 5. 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 **20.10.2024 23:59**.

Fáza 2 - Predspracovanie údajov: 15 bodov

Správa sa odovzdáva v 7. 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 **03.11.2024 23:59**.

Fáza 3 - Strojové učenie: 20 bodov

Správa sa odovzdáva v 10. 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 **24.11.2024 23:59**.

Aktivity na cvičení : 10 bodov

Správa sa odovzdáva v 12. 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 **08.12.2024 23:59**.