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 projektu

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 riešite 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

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 niečo treba 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 techniky znovu nebudeme. Zmeny zvolených postupov dokumentujte. Problém s dátami môžete riešiť iteratívne v každej fáze a vo všetkých fázach podľa potreby.

2.1 Realizácia predspracovania dát (5b).

- (A-1b) Dáta si rozdeľte na trénovaciu a testovaciu množinu podľa vami preddefinovaného pomeru. Ďalej pracujte len **s trénovacím datasetom**.
- (B-1b) Transformujte dáta na vhodný formát pre ML t.j. jedno pozorovanie musí byť opísané jedným riadkom a každý atribút musí byť v numerickom formáte (encoding). zIteratívne integrujte aj kroky v predspracovaní dát z prvej fázy (missing values, outlier detection) ako celok.
- (C-2b) Transformujte atribúty dát pre strojové učenie podľa dostupných techník minimálne: scaling (2 techniky), transformers (2 techniky) a ďalšie. Cieľom je aby ste testovali efekty a vhodne kombinovali v dátovom pipeline (od časti 2.3 a v 3. fáze).
- (D-1b) Zdôvodnite Vaše voľby/rozhodnutie pre realizáciu (t.j. zdokumentovanie)

2.2 Výber atribútov pre strojové učenie (5b)

- (A-3b) Zistite, ktoré atribúty (features) vo vašich dátach pre ML sú informatívne k predikovanej premennej (minimálne 3 techniky s porovnaním medzi sebou).
- (B-1b) Zoraďte zistené atribúty v poradí podľa dôležitosti.
- (C-1b) Zdôvodnite Vaše voľby/rozhodnutie pre realizáciu (t.j. zdokumentovanie)

2.3 Replikovateľnosť predspracovania (5b)

- (A-3b) Upravte váš kód realizujúci predspracovanie trénovacej množiny tak, aby ho bolo možné bez ďalších úprav znovu použiť **na predspracovanie testovacej množiny** v kontexte strojového učenia.
- (B-2b) Využite možnosti **sklearn.pipeline**

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

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**.

3.1 Jednoduchý klasifikátor na základe závislosti v dátach (5b)

- (A-3b) Naimplementujte jednoduchý **ID3** klasifikátor s hĺbkou min 2 (vrátane root/koreň).
- (B-1b) Vyhodnoťte Váš ID3 klasifikátor pomocou metrík accuracy, precision a recall.
- (C-1b) Zístite či Váš ID3 klasifikátor má overfit.

3.2 Trénovanie a vyhodnotenie klasifikátorov strojového učenia (5b)

- (A-1b) Na trénovanie využite jeden **stromový algoritmus** v scikit-learn.
- (B-1b) Porovnajte s jedným iným **nestromovým algoritmom** v scikit-learn.
- (C-1b) Porovnajte výsledky s ID3 z prvého kroku.
- (D-1b) Vizualizujte natrénované pravidlá **minimálne** pre jeden Vami vybraný algoritmus
- (E-1b) Vyhodnoťte natrénované modely pomocou metrík accuracy, precision a recall

3.3 Optimalizácia alias hyperparameter tuning (5b)

- (A-1b) Vyskúšajte rôzne nastavenie hyperparametrov (tuning) pre zvolený algoritmus tak, aby ste optimalizovali výkonnosť (bez **underfitting**u**)**.
- (B-1b) Vyskúšajte kombinácie modelov (ensemble) pre zvolený algoritmus tak, aby ste optimalizovali výkonnosť (bez **underfitting**u**)** .
- (C-1b) Využite krížovú validáciu (**cross validation)** na trénovacej množine.
- (D-2b) Dokážte že Váš nastavený najlepší model je bez **overfitting**u**.**

3.4 Vyhodnotenie vplyvu zvolenej stratégie riešenia na klasifikáciu (5b)

Vyhodnoťte Vami zvolené stratégie riešenia projektu z hľadiska classification accuracy, či sú učinné pre Váš dataset:

- (A-1b) Stratégie riešenia chýbajúcich hodnôt a outlierov
- (B-1b) Dátová transformácia (scaling, transformer, ...)
- (C-1b) Výber atribútov, výber algoritmov, hyperparameter tuning, ensemble learning
- (D-1b) Ktorý model je Váš **najlepší model** pre nasadenie (deployment)?
- (E-1b) 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 tie metadáta sú potrebné a vyrobené v developmente.

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

The QUEST (stačí jeden z dvoch, buď Q1 alebo Q2)

- Q1 (Image classification): klasifikačná úloha podľa počtu tried
- Q2 (Time-series forecasting): predpovedajte čo najpresnejšie nastávajúcu situáciu podľa historických dát.

Nemusíte použiť celý dataset, len toľko koľko stačí na modelovanie s zdôvodnením že máte dostatok dát.

Vyberte si len jeden z datasetov na riešenie "*the quest*" podľa bloku cvičenia
Alqnatri

- Q1 Age Detection Dataset 50 MB, Age recognition, 3 classes: old, middle, young
- Q2 Household Electric Power Consumption 131 MB, Time-Series Forecasting

Bakonyi

- Q1 Periocular recognition 14 MB, Facial recognition with medical mask, 2 classes
- Q2 S&P 500 Stocks 202 MB, Time-Series Forecasting

Kollár

- Q1 Covid-19 Image Dataset 166 MB, 3 classes: COVID-19, Viral Pneumonia, Norma
- Q2 Household Electric Power Consumption 131 MB, Time-Series Forecasting

Lytvyn

- Q1 Head CT - hemorrhage data 26 MB, Tumor detection, 2 classes: normal, hemorrhage
- Q2 Netflix 10+ Year Stock Data 303 KB, Time-Series Forecasting

Nguyen

- Q1 Covid-19 Image Dataset 166 MB, 3 classes: COVID-19, Viral Pneumonia, Norma
- Q2 S&P 500 Stocks 202 MB, Time-Series Forecasting

4.1 EDA and data preprocessing (5b)

- (A-4b) EDA a data preprocessing pre Vami vybrané charakteristiky z datasetu
- (B-1b) Zdôvodnite výber ML/DL metód vzhľadom na Vami vybraný dataset pre 4.2

4.2. Modeling and evaluation (5b)

- (A-4b) Modeluje Vami tie vybrané charakteristiky pomocou vhodných ML/DL metód. Výsledok modelovania je najlepší model.
- (B-1b) Zhodnotíte Váš prístup a získaný výsledok

Všetky hodnotenia podložte dôkazmi.
Najlepší model má byť stabilný, bez overfitu a bez underfitu.

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**.