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.

Dáta

https://drive.google.com/drive/folders/1DJbqyN_TtSfHdg4_qQl6dbqsVmLGlWO1?usp=sharing
(každá dvojica má jeden dataset pod číslom, ktoré máte na cvičení)

Doplňujúce informácie o dátach: vysvetlenie vybraných senzorových hodnôt

Saturácia kyslíkom (angl. **oxygen saturation**) je kľúčovým ukazovateľom správneho fungovania dýchacej a obehovej sústavy. Ak jej hodnota klesne na kriticky nízku úroveň, môže to signalizovať život ohrozujúce stavy, ako sú hypoxémia, respiračné zlyhanie alebo závažné infekcie. V takýchto prípadoch je nevyhnutný rýchly zásah. Tradičné monitorovanie sa realizuje pomocou pulzných oxymetrov, ktoré však môžu byť ovplyvnené šumom, pohybovými artefaktmi alebo majú obmedzenia v niektorých klinických situáciách.

Moderné prístupy založené na strojovom učení prinášajú možnosť presnejšie odhadovať a predikovať kritické hodnoty saturácie kyslíkom (**critical oxygen saturation estimation**). Modely môžu využívať multimodálne údaje, ako sú srdcová frekvencia, dychová frekvencia, krvný tlak či signály zo senzorov. Vďaka trénovaniu na rôznorodých dátach je možné identifikovať skoré varovné príznaky desaturácie, odfiltrovať šum a poskytnúť včasné upozornenia ešte pred poklesom saturácie pod bezpečnú hranicu.

Cieľom tohto zadania je oboznámiť sa s problematikou monitorovania saturácie kyslíkom, pochopiť prínos umelej inteligencie a navrhnúť riešenie, ktoré by mohlo prispieť k zlepšeniu kritickej starostlivosti a zníženiu rizík spojených s neodhalenou hypoxémiou.

Zadanie projektu

Critical oxygen saturation estimation

The QUEST

Každá dvojica bude pracovať s pridelenou dátovou sadou od 2. týždňa. **Vašou úlohou** je predikovať závislé hodnoty premennej "**oximetry**" **(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 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 a či spĺňa predpísané podmienky a rozsah meraných hodnôt.
- (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, ktoré môžu vznikať po určitých dátových transformáciach, nejednotné formáty, chýbajúce hodnoty, vychýlené hodnoty. V dátach sa môžu nachádzať aj iné, tu nevymenované problémy, resp. menej problémov ako bolo uvedených.
- (B-2b) Kontrola správnosť v dátach
  - či obsahujú abnormálne hodnoty
  - či obsahujú nelogické dátové vzťahy, ktoré sú následkom dátovej kolekcie a anotovania dát
- (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:

(SK) FiO_2 má v priemere vyššiu hodnotu v stave s oximetry ako bez nej.
(EN) FiO_2 has, on average, a higher value in the state with oximetry than without it.

- (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 **26.10.2025 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 strojové učenie t.j. jedno pozorovanie musí byť opísané jedným riadkom a každý atribút musí byť v numerickom formáte. Iteratívne integrujte aj kroky v predspracovaní dát z prvej fázy ako celok.
- (C-1b) 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-2b) 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 **09.11.2025 23:59**.

Fáza 3 - Strojové učenie: 20 bodov

Správa sa odovzdáva v 10. týždni semestra. Dvojica svojmu cvičiacemu odprezentuje vykonanú fázu v Jupyter Notebooku na cvičení. V notebooku uveďte percentuálny podiel práce členov dvojice. Následne správu elektronicky odovzdá **jeden člen z dvojice** do systému **AIS** do nedele **30.11.2025 23:59**.

Aktivity na cvičení: 10 bodov

Správa sa odovzdáva v 12. týždni semestra. Dvojica svojmu cvičiacemu odprezentuje vykonanú fázu v Jupyter Notebooku na cvičení. V notebooku uveďte percentuálny podiel práce členov dvojice. Následne správu elektronicky odovzdá **jeden člen z dvojice** do systému **AIS** do nedele **14.12.2025 23:59**.