

# 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**, atď.. 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\\_qQl6dbqsVmLGIWO1?usp=sharing](https://drive.google.com/drive/folders/1DJbqyN_TtSfHdg4_qQl6dbqsVmLGIWO1?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 premenná**) 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.

Očakávaným **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 a či spĺňa predpísané podmienky a rozsah meraných hodnôt.
- (C-1b) Párová analýza dát: Identifikujte vzťahy a závislosti 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 vzniknúť 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 odlahľých 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:  
**FiO<sub>2</sub> má v priemere vyššiu hodnotu v stave s oximetry ako bez nej.**
- (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

**Správa sa odovzdáva v 7. týždni semestra.** Dvojica svojmu cvičiacemu odprezentuje vykonanú fázu v notebooku podľa potreby na cvičení. Uvedte 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 uvedte 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 uvedte percentuálny podiel práce členov dvojice. Následne správu elektronicky odovzdá **jeden člen z dvojice** do systému **AIS** do nedele **14.12.2025 23:59**.