# 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 dostupných knižníc pre dátovú vedu ako **pandas, numpy, scipy, 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 sa predstavuje významnú zložku hodnotenia.

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

# Dáta

<https://drive.google.com/drive/folders/1wZaVpr0VedXeS1TgjGOg6eHkhwvWA2BM?usp=sharing>

Znečistenie ovzdušia spôsobuje vážne dýchacie a srdcové ochorenia, ktoré môžu byť smrteľné. Najčastejšie sú postihnuté deti, čo vedie k zápalu pľúc a problémom s dýchaním vrátane astmy. Kyslé dažde, ničenie ozónovej vrstvy a globálne otepľovanie sú niektoré z nepriaznivých dôsledkov. Dátová sada pre Vás (World's Air Pollution: Real-time Air Quality Index <https://waqi.info/>) predstavuje záznamy jednotlivých meraní kvality ovzdušia ako kombinácia mnohých faktorov bez časovej následnosti. V záznamoch je závislá premenná s menom **“*warning*”** indikujúca alarmujúci stav kvality ovzdušia. Vo veľkých mestách ako napr. Peking (angl. Beijing, hlavné mesto Číny s viac ako 21 miliónov ľudí) sa pri varovaní spustí opatrenie ako obmedzenie pohybov áut a ľudí v meste alebo umelý dážď až pokiaľ kvalita vzduchu sa nevráti do normu.

**Slovník odborných skratiek v doméne, ktoré sa vyskytujú v datách**

PM2.5 Particulate Matter (µg/m3)

PM10 Particulate Matter (µg/m3)

NOx Nitrogen Oxides (µg/m3)

NO2 Nitrogen Dioxide (µg/m3)

SO2 Sulfur Dioxide (µg/m3)

CO Carbon Monoxide emissions (µg/m3)

CO2 Carbon Dioxide (µg/m3)

PAHs Polycyclic Aromatic Hydrocarbons (µg/m3)

NH3 Ammonia trace (µg/m3)

Pb Lead (µg/m3)

TEMP Temperature (degree Celsius)

DEWP Dew point temperature (degree Celsius)

PRES Pressure (hPa, <100, 1050>)

RAIN Rain (mm)

WSPM Wind Speed (m/s)

WD Wind Direction

VOC Volatile Organic Compounds

CFCs Chlorofluorocarbons

C2H3NO5 Peroxyacetyl nitrate

H2CO Plywood emit formaldehyde

GSTM1 Glutathione-S transferase M1

1-OHP 1-hydroxypyrene

2-OHF 2-hydroxyfluorene

2-OHNa 2-hydroxynaphthalene

N2 Nitrogen

O2 Oxygen

O3 Ozone

Ar Argon

Ne Neon

CH4 Methane

He Helium

Kr Krypton

I2 Iodine

H2 Hydrogen

Xe Xenon

Vybrané stĺpce obsahujú škálované resp. spriemernené hodnoty z rôznych časových intervalov. Dôvod je aplikovanie rôznych štandardov platných v rôznych krajinách sveta.

# Zadanie

Každá dvojica bude pracovať s pridelenou dátovou sadou od 3. týždňa.

* Vašou úlohou je predikovať závislé hodnoty premennej “***warning***” pomocou metód strojového učenia

Budete sa musieť pritom vysporiadať s viacerými problémami, ktoré sa v dátach nachádzajú ako formáty dát, chýbajúce, vychýlené hodnoty a pod.

### 

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

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

V tejto fáze uveďte:

* Počet záznamov, počet atribútov, ich typy,
* Pre zvolené významné atribúty ich distribúcie, základné deskriptívne štatistiky a pod.
* Párová analýza dát: preskúmajte vzťahy medzi zvolenými dvojicami atribútov.
* Párová analýza dát: Identifikujte závislostí medzi dvojicami atribútov (napr. korelácie)
* Párová analýza dát: Identifikujte závislosti medzi predikovanou premennou a ostatnými premennými (potenciálnymi prediktormi).

#### Identifikácia problémov v dátach s prvotným riešením (5b)

* Identifikujte problémy v dátach napr.: nevhodná štruktúra dát, duplicitné záznamy, nejednotné formáty, chýbajúce hodnoty, vychýlené hodnoty. V dátach sa môžu nachádzať aj iné, tu nevymenované problémy.

#### 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)

* Sformulujte **dve hypotézy** o dátach v kontexte zadanej predikčnej úlohy.

Príklad formulovania hypotézy: *merania kvality ovzdušia v kritickom stave majú v priemere inú (vyššiu/nižšiu) hodnotu určitej chemikálie (alebo koncentrácie látok) ako merania kvality ovzdušia v normálnom stave*.

* Sformulované hypotézy overte vhodne zvoleným štatistickým testom.

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

1. Majú dáta vhodný formát pre ďalšie spracovanie? Ak nie, aké problémy sa v nich vyskytujú?
2. Sú niektoré atribúty medzi sebou závislé? Od ktorých atribútov závisí predikovaná premenná?
3. Sú v dátach chýbajúce hodnoty? Ako plánujete riešiť tento problém?
4. Nadobúdajú niektoré atribúty nekonzistentné alebo výrazne odchýlené hodnoty?
5. Ako plánujete/riešíte tieto identifikované problémy?

Správa sa odovzdáva v 6. týždni semestra na cvičení. Dvojica svojmu cvičiacemu odprezentuje vykonanú prieskumnú analýzu v Jupyter Notebooku). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do nedele **30.10.2022 23:59**.

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

Správa sa odovzdáva v 9. týždni semestra na cvičení. Dvojica svojmu cvičiacemu odprezentuje vykonané predspracovanie v Jupyter Notebooku). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do nedele 20.11.2022 23:59.

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

Správa sa odovzdáva v 12. týždni semestra na cvičení. Dvojica svojmu cvičiacemu odprezentuje výsledky strojového učenia v Jupyter Notebooku). Následne správu elektronicky odovzdá jeden člen z dvojice do systému AIS do nedele 11.12.2022 23:59.