# 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, statsmodels, 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 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**. 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 **30.10.2022 23:59**.

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

V tejto fáze sa od Vás očakáva že realizujte **predspracovanie údajov** pre strojové učenie. Výsledkom bude upravená dátová sada (csv alebo tsv), kde jedno pozorovanie je opísané jedným riadkom.

* **scikit-learn** vie len numerické dáta, takže treba niečo 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 možné že treba realizovať EDA opakovane podľa Vašej potreby. Bodovanie znovu (EDA) nebudeme, zmeny ale dokumentujte. Problém s dátami môžete riešiť iteratívne v každej fáze aj vo všetkých fázach podľa potreby.

#### Integrácia a čistenie dát (5b)

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.

* Chýbajúce hodnotu (missing values): vyskúšajte min. 2 techniky ako napr.
  + odstránenie pozorovaní s chýbajúcimi údajmi
  + nahradenie chýbajúcej hodnoty mediánom, priemerom, pomerom (ku korelovanému atribútu), alebo pomocou lineárnej regresie resp. kNN
* Podobne postupujte aj pri riešení vychýlených hodnôt (outlier detection), min. 2 techniky:
  + odstránenie vychýlených (odľahlých) pozorovaní
  + nahradenie vychýlenej hodnoty hraničnými hodnotami rozdelenia (napr. 5%, 95%)

#### Realizácia predspracovania dát (5b).

* Transformované dáta pre strojové učenie si rozdeľuje na trénovaciu a testovaciu množinu podľa vami preddefinovaným pomerom. Naďalej pracujte len **s trénovacím datasetom**.
* Transformujte atribútov dát pre strojové učenie podľa dostupných techník (minimálne 2 techniky) ako scaling, transformers a ďalšie.
* Zdôvodnite Vašu voľby/rozhodnutie pre realizáciu (t.j. zdokumentovanie)

#### Výber atribútov pre strojové učenie (5b)

* Zistite ktoré atribúty (features) vo vašich dátach pre strojové učenie sú informatívne k atribútu “***warning***”. Zoradíte tie atribúty v poradí podľa dôležitosti.
* Zdôvodnite Vašu voľby/rozhodnutie pre realizáciu (t.j. zdokumentovanie)

#### Replikovateľnosť predspracovania (5b)

* Upravte váš kód realizujúci predspracovanie trénovacej množiny tak, aby ho bolo možné bez ďalších úprav znovu použiť **na predspracovanie testovacej množiny** (pomocou funkcie/í)
* Očakáva sa aj využitie možnosti **sklearn.pipeline**

**Správa sa odovzdáva v 9. 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 **20.11.2022 23:59**.

## Fáza 3 - Strojové učenie (v 12. týždni): 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**.

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

* Naimplementujte OneR algorithm (iné mená: OneRule or 1R), ktorý je jednoduchý klasifikátor tzv. rozhodnutie na základe jedného atribútu. Môžete implementovať aj komplikovanejšie t.j. rozhodnutie na základe kombinácie atribútov.
* Algoritmus by mal byť realizovaný na základe závislostí v dátach. Vyhodnoťte klasifikátora pomocou metrík accuracy, precision a recall.

#### Trénovanie a vyhodnotenie klasifikátorov strojového učenia (5b)

* Na trénovanie využite **minimálne jeden stromový algoritmus** strojového učenia v scikit-learn.
* Vizualizujte natrénované pravidlá.
* Vyhodnoťte natrénované modely pomocou metrík accuracy, precision a recall
* Porovnajte ašpoň jeden natrénovaný klasifikátor v scikit-learn s jednoduchým klasifikátorom z prvého kroku.

#### Optimalizácia alias hyperparameter tuning (5b)

* Preskúmajte hyperparametre Vášho zvoleného klasifikačného algoritmu v druhom kroku a vyskúšajte ich rôzne nastavenie tak, aby ste **minimalizovali overfitting** (preučenie) a **optimalizovali** výsledok.
* Vysvetlite, čo jednotlivé hyperparametre robia. Pri nastavovaní hyperparametrov algoritmu využite **krížovú validáciu** (cross validation) na trénovacej množine.

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

Vyhodnotíte Vami zvolené stratégie riešenia projektu z hľadiska classification accuracy:

* Stratégie riešenia chýbajúcich hodnôt a outlierov;
* Scaling resp. transformer či zlepší accuracy klasifikácie;
* Výber atribútov a výber algoritmov strojového učenia;
* Hyperparameter tuning resp. ensemble learning.

Ktoré spôsoby z hore-uvedených bodov sa ukázali ako učinné pre Váš dataset? Hodnotenie podložíte dôkazmi.

**Správa sa odovzdáva v poslednom týždni semestra.** Dvojica svojmu cvičiacemu odprezentuje vykonanú fázu v Jupyter Notebooku podľa potreby na cvičení. V notebooku uveďte percentuálny podiel prácečlenov dvojice. Následne správu elektronicky odovzdá **jeden člen z dvojice** do systému **AIS** do štvrtka **15.12.2022 23:59**.