**Uczenie maszynowe**

Zasady ćwiczeń:

1. Format rozwiązania: notebook (.ipynb). Widzimy kod, rozwiązanie, a jak trzeba grafikę.
2. Wykonane skrypty wkładamy do Zadania na platformie Moodle (czas jest ograniczony).
3. Na zajęciach należy zreferować rozwiązania (proszę, żeby nie było sytuacji, że nikt nie referuje).

Lab: **4** Analiza/rozpoznanie/przygotowanie danych.

Jeśli metodykę któregoś z etapów pracy można uogólnić, taki etap proszę zakodować w postaci metody  
(z ew. parametrami).

1. Proszę pobrać dowolny zbiór danych (każdy Student inny).

Przykładowe archiwa danych:

*www.kaggle.com/datasets*

*archive.ics.uci.edu/ml*

*datasetsearch.research.google.com*

1. Poznać motywację badawczą zbioru danych: dziedzina, problem, rodzaj uczenia.
2. Zwizualizować rozkłady cech, w tym celu wykorzystać histogramy, rozkłady rozrzutu, lub inne wykresy.
3. Określić wartości odstające (outliers).
4. Dla każdej cechy zebrać statystyki: średnia, wariancja, wartość najmniejsza/największa.
5. Określić kowariancję/korelację cech.
6. Przeanalizować, które cechy są potencjalnie najlepsze do uczenia.
7. W oparciu o powyższe wybrać do uczenia: określone klasy przypadków, wybrane cechy, oraz liczbę rekordów danych (wszystkie?). Przygotowany zbiór zapisać.
8. Proszę pobrać zbiór danych, w którym są wartości brakujące (NaN/0/-/(puste miejsce)) lub usunąć ze zbioru danych losowo pewną liczbę komórek.
   1. Wyznaczyć statystykę braków i ocenić jak ten problem rozwiązać.
   2. Obsłużyć wartości brakujące za pomocą usunięcia rekordów.
   3. Obsłużyć wartości brakujące wypełniając je zerem/inną wartością stałą.
   4. Obsłużyć wartości brakujące wypełniając wartością średnią cechy (wartości ciągłe) lub medianą (wartości dyskretne).
9. Przeprowadzić
   1. skalowanie każdej cechy w zadanym przedziale (domyślnie [-1,1]),
   2. normalizację każdej cechy (do rozkładu Gaussa N(0,1)).
10. Zakodować etykiety klas
    1. problemy binarne: 0/1, -1/1,
    2. problemy wieloklasowe: kolejne cyfry, one-hot encoding.