# AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

**„Uczenie maszynowe" – laboratorium**

# Laboratorium 2

# Klasyfikacja algorytmem k-najbliższych sąsiadów (k-nn)

data aktualizacji: 20.03.2023

## Cel ćwiczenia

Celem ćwiczenia laboratoryjnego jest uruchomienie algorytmu klasyfikacji k-nn, zbadanie jego skuteczności klasyfikacji dla trzech zbiorów testowych, analiza jak radzi sobie przy różnych zbiorach danych oraz sprawdzenie jak wartości parametrów algorytmu k-nn wpływają na uzyskiwane wyniki. W tym celu użyte będą klasyczne miary jakości klasyfikacji.

Dodatkowo, w ramach ćwiczenia użyta będzie procedura walidacyjna, oraz jej modyfikacja dla zbalansowania zbiorów treningowych/testowych.

## Sugerowane narzędzia

Zadanie jest realizowane przy pomocy języka python i bibliotek użytecznych przy manipulacji danymi, klasyfikacji (k-nn), oceny jakości klasyfikatora, walidacji krzyżowej (również stratyfikowanej) oraz wizualizacji wyników (tabele/wykresy).

Sugerowane narzędzia (python): jupyter, sklearn, scipy, pandas, matplotlib, seaborn itp.

## Użyte zbiory danych

W ćwiczeniu użyte będą powszechnie używane zbiory:

- IRIS – https://archive.ics.uci.edu/ml/datasets/iris

- WINE – https://archive.ics.uci.edu/ml/datasets/wine
- Polish Companies Bankruptcy – https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

## Przebieg ćwiczenia

1. Uruchomienie algorytmu klasyfikacji **k-nn** dla zbioru IRIS – używamy wartości parametrów domyślnych (np. k=5, odległość Euklidesa, głosowanie większościowe)
2. Przyswojenie wiadomości z wykładu dotyczące zasady budowy macierzy pomyłek (ang. *confusion matrix*), oraz interpretacji (oceny) wyniku klasyfikatora dla instancji (danych), takich jak TP, TN, FN, FP. Na postawie tych składowych następuje implementacja miar **jakości klasyfikacji**: Precision, Recall, F-Score i Accuracy.

Uwaga! Jeśli chcemy używać jednej miary, sugeruje się użycie **F-score** (F-measure), która jest złożeniem miar Precision i Recall.

3. Zbadanie jaka jest skuteczność klasyfikacji k-nn dla zbioru IRIS.
4. Implementacja procedury **walidacji krzyżowej** (ang. *crossvalidation*) dla różnych rozmiarów, tj. 2-fold, 5-fold, 10-fold i sprawdzenie jak wpływa na wyniki klasyfikatora k-nn dla zbioru IRIS. Jak liczba foldów wpływa na skuteczność?

**Pytanie pomocnicze nr 1:** czy powinno używać się procedury mieszania danych (tj. *shuffle*)?

**Pytanie pomocnicze nr 2:** czy warto rozważyć użycie walidacji krzyżowej typu *leave-one-out*?

5. Sprawdzenie różnych **wartości parametrów** klasyfikatora k-nn dla zbioru IRIS:
   - liczba *k*, która określa liczbę sąsiadów. Warto sprawdzić wiele wartości, np. z przedziału 1-15. Warto także zwrócić uwagę czy parzysta i nieparzysta liczba sąsiadów daje podobne wyniki? Dlaczego?
   - 3 sposoby głosowania: większościowe, ważone (np. odległością), oraz inne (zaproponowane przez studenta)

- 3 miary odległości w przestrzeni danych – warto sprawdzić nie tylko odległość Euklidesa, manhattańską, ale także miarę Minkowskiego, Mahalanobisa czy Czebyszewa

6. Zbadać wg kroków 3-5 skuteczność dla zbioru WINE
7. Zbadać wg kroków 3-5 skuteczność dla zbioru PCB
8. Zbadać jak/czy zmienia się skuteczność klasyfikacji k-nn dla badanych zbiorów jeśli zastosowana zostanie walidacja stratyfikowana? Dlaczego? Przy realizacji zadania należy rozważyć: czy dane powinny być powtórzone? Czy uśredniać wyniki dla miar, foldów?

Przy realizacji zadania należy użyć tabel oraz odpowiednich typów wykresów (np. liniowy, słupkowy, skrzynkowy).

## Punktacja

Przy realizacji zadania można otrzymać **max 8 punktów** wedle poniższej punktacji:

| | |
|---|---|
| 1 | Uruchomienie k-nn dla trzech badanych zbiorów |
| 2 | Sprawdzenie skuteczności (F-score) klasyfikacji dla różnych zestawów wartości parametrów dla 3 różnych zbiorów |
| 2 | Sprawdzenie skuteczności dla różnych typów głosowania (zbiory WINE i PCB) |
| 2 | Sprawdzenie jak na skuteczność wpływa liczba *k* (zbiory WINE i PCB) |
| 1 | Sprawdzenie jak na skuteczność wpływa liczba foldów w walidacji krzyżowej oraz jej typ (zwykła/stratyfikowana) (zbiory WINE i PCB) |

Przy realizacji tego zadania wystarczy prosty raport PDF utworzony przy użyciu Jupyter, w którym będą tabelki/wykresy, też wnioski wyciągnięte ze zrealizowanego ćwiczenia.

Przy prezentowaniu skuteczności (np. F-score) i używaniu walidacji krzyżowej możemy posługiwać wartością średnią (ale też min, max i standardowe odchylenie) dla danego zestawu foldów. Aby zaprezentować (i porównać) skuteczność klasyfikacji dla różnych zestawów parametrów sugeruje się użycie **wykresów pudełkowych** (ang. boxplot), które pozwalają na kompleksową prezentację w/w wartości.

Przy wynikach badań należy skomentować, podać wnioski i podsumowanie.

## Pytania pomocnicze

1. Co jest modelem klasyfikacji w tym zadaniu? Dlaczego k-nn jest typu *lazy learning*?
2. Ile parametrów ma klasyfikator k-nn? Które z nich są istotne dla skuteczności?
3. Czym się różni kroswalidacja stratyfikowana od „zwykłej”?
4. Dlaczego zwykle nie stosujemy walidacji krzyżowej leave-one-out?
5. Czy k-nn wymaga standaryzacji/normalizacji danych?
6. Dlaczego miara Accuracy dla zbiorów niezbalansowanych jest mniej użyteczna niż miara F-score?

## Literatura

1. Materiały do wykładu

2. Cichosz P. ”Systemy uczące się”, WNT Warszawa

3. Zasoby Internetu, słowa kluczowe: uczenie maszynowe (machine learning), data mining, crossvalidation, stratified crossvalidation, data normalisation, data standarisation, confusion matrix, precision, recall, f-score, accuracy, k-nearest neighbors