AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe" – laboratorium

Laboratorium 0

Data aktualizacji: 03.03.2024

Wprowadzenie do *Python*

Cel ćwiczenia

Celem ćwiczenia laboratoryjnego jest uruchomienie (wraz z instalacją) środowiska programistycznego języka Python oraz narzędzi potrzebnych do realizacji zadań następnych list. W trakcie realizacji zadania wczytane zostaną standardowe zbiory danych, które będą podstawą dokładniejszej analizy. Użyty zostanie algorytm PCA i biblioteki wizualizacji danych.

Dostępność materiałów i narzędzi

Narzędzia oraz ich dokumentacja jest ogólnodostępna w sieci Internet na licencji *opensource*.

Sugerowane narzędzia

- Python w wersji 3.x – jako język i środowisko oprogramowania – https://www.python.org/
- Jupyter (notebook) – środowisko programowania/generowania dokumentacji – https://jupyter.org/
- scikit learn – biblioteka python modeli do uczenia maszynowego – https://scikit-learn.org/stable/
- scipy – zbiór bibliotek python do operacji na danych – https://www.scipy.org/ Szczególnie przydatne:
    - pandas – struktury danych i analizy – https://pandas.pydata.org/

    - numpy – przydatna biblioteka do obliczeń w python – https://numpy.org/
    - mathplotlib – biblioteka do wizualizacji (wykresy) w python – https://matplotlib.org/stable/
- seaborn – zaawansowana biblioteka wizualizacji danych – https://seaborn.pydata.org/
- plotly – zaawansowana biblioteka wizualizacji danych - https://plotly.com/python/

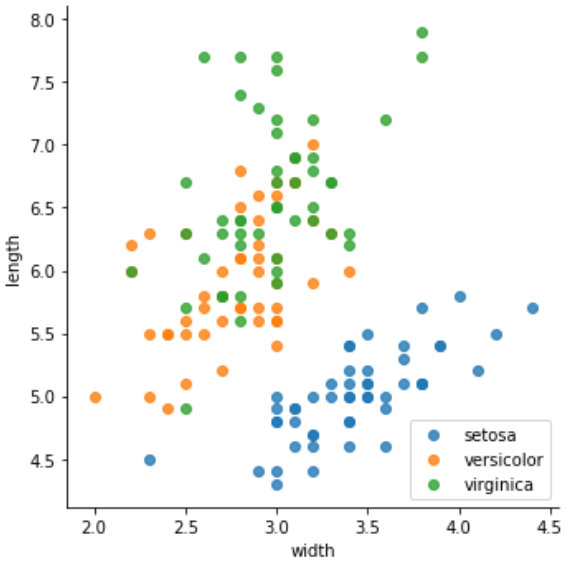
Zbiory danych

W ćwiczeniu użyte będą powszechnie używane zbiory:

- IRIS – https://archive.ics.uci.edu/ml/datasets/iris
- GLASS – https://archive.ics.uci.edu/ml/datasets/glass+identification
- WINE - https://archive.ics.uci.edu/ml/datasets/wine

Przebieg ćwiczenia

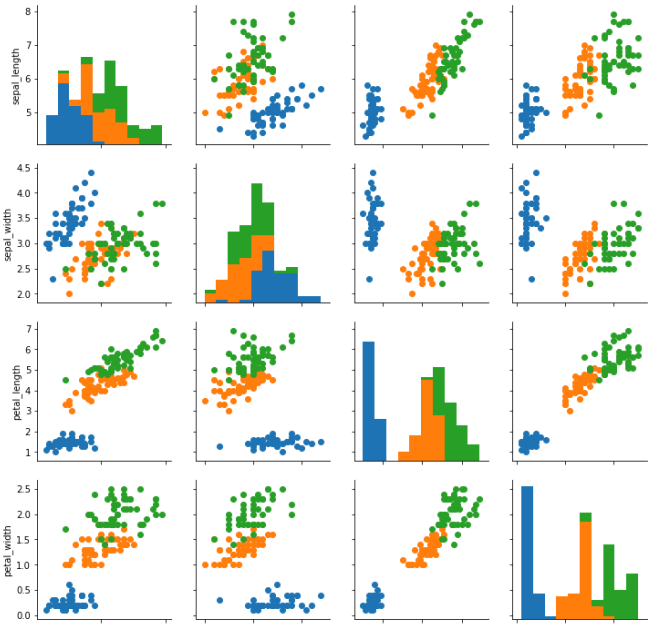
1. Instalacja Python wraz z niezbędnymi bibliotekami.
2. Instalacja środowiska programistycznego (np. Jupyter)
3. Wczytanie zbioru IRIS, WINE, GLASS
4. Statystyczna analiza zbiorów IRIS, WINE, GLASS, np. klasy (liczba, interpretacja), instancje, atrybuty, dystrybucja klas w zbiorze.
5. Wizualizacja oraz analiza zbiorów IRIS, WINE, GLASS, np. Wykres 1, Wykres 2.



Wykres 1. Zależność długość i szerokości kielicha w zbiorze danych IRIS.

6. Użycie algorytmu PCA, wizualizacja oraz analiza wyników.

   Algorytm PCA (ang. *principal component analysis*) tj. wyznaczania głównych składowych analizowanego zbioru. PCA stosuje się do zmniejszenia wymiarowości zbioru (więcej informacji w literaturze poniżej).



Wykres 2. Zależności zmiennych w zbiorze danych IRIS.

Punktacja

Przy realizacji zadania student może otrzymać **max 5 punktów** wedle poniższej punktacji.

1	Instalacja środowiska z niezbędnymi bibliotekami
1	Wczytanie zbioru IRIS, wyrysowanie wykresu zależności długości/szerokości płatków (jak Wykres 1), Analiza zbioru i wizualizacja rozkładu danych.
1	Wczytanie zbioru GLASS, wyrysowanie wykresu zależności wybranych atrybutów (jak Wykres 1), Analiza zbioru i wizualizacja rozkładu danych.
1	Wczytanie zbioru WINE, wyrysowanie wykresu zależności wybranych atrybutów (jak Wykres 1), Analiza zbioru i wizualizacja rozkładu danych.
1	Użycie PCA i narysowanie wykresu wynikowego dla trzech zbiorów

Pytania pomocnicze

1. Czym się różnią zbiory danych analizowane w treści zadania? Na czym może polegać „trudność" analizy. Który z nich wydaje się być łatwiejszy/trudniejszy?
2. Czy nierównomierny rozkład klas w zbiorze może stanowić problem dla analizy i dalszej budowy modelu danych?
3. Jak działa PCA i kiedy warto go stosować?

Literatura

1. https://scikit-learn.org/stable/modules/decomposition.html#pca
2. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA