2

PRZETWARZANIE DANYCH

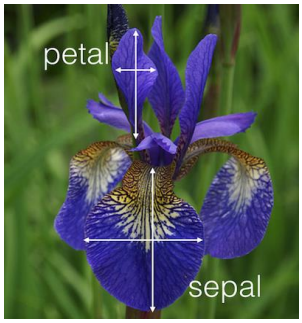
SZUKANIE BŁĘDÓW, BRAKÓW; PCA; NORMALIZACJA

DATASET Z IRYSAMI

Dataset z irysami (iris dataset) to jeden z najbardziej popularnych zbiorów danych do prostych testów. Zawiera 150 rekordów z pomiarami płatków trzech gatunków irysa:

- Iris setosa
- Iris virginica
- Iris versicolor

Po 50 z każdego gatunku. Dla każdego irysa wykonano pomiar długości i szerokości płatka wywiniętego do góry (petal) i płatka wywiniętego w dół (sepal). Dodatkowo w piątej kolumnie datasetu są informacje o gatunku.
Na tych laboratoriach będziemy działali na tym zbiorze danych.



	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Do przetwarzania pliku csv zzawierającego dane warto użyć paczki pandas. Przetestuj:

```
import pandas as pd

df = pd.read_csv("iris.csv")

print(df)
```

W obiekcie `df` mamy bazę danych typu `dataframe`. Żeby bezpośrednio wejść do danych z tej bazy trzeba odwołać się do wartości (values).

```
print(df.values)
```

Spróbujmy wyświetlić różne fragmenty danych.

```
#wszystkie wiersze, kolumna nr 0
print(df.values[:, 0])

#wiersze od 5 do 10, wszystkie kolumny
print(df.values[5:11, :])

#dane w komórce [1,4]
print(df.values[1, 4])
```


Grzegorz Madejski
grzegorz.madejski@ug.edu.pl
Zakład Sztucznej Inteligencji,
Instytut Informatyki, Uniwersytet Gdański

💻 ZADANIE 1: **BŁĘDY I BRAKUJĄCE DANE W IRYSACH**

Ściągnij ze strony zajęć plik iris_with_errors.csv. Znajduje się w nim baza danych z irysami, jednak są w niej błędy. Celem zadania jest poprawienie tych błędów. Zamiast jednak ręcznego szukania i poprawiania, wykorzystajmy do tego specjalistyczne paczki. W języku Python można to zrobić z wykorzystaniem paczki Pandas:

- https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b
- https://realpython.com/python-data-cleaning-numpy-pandas/

Wykorzystując powyższe linki i technologie w nich zaprezentowane napraw bazę iris_with_errors.csv. Postaraj się wykonać następujące kroki.

a) Policz ile jest w bazie brakujących lub nieuzupełnionych danych. Wyświetl statystyki bazy danych z błędami.
b) Sprawdź czy wszystkie dane numeryczne są z zakresu (0; 15). Dane spoza zakresu muszą być poprawione. Możesz tutaj użyć metody: za błędne dane podstaw średnią (lub medianę) z danej kolumny.
c) Sprawdź czy wszystkie gatunki są napisami: „Setosa", „Versicolor" lub „Virginica". Jeśli nie, wskaż jakie popełniono błędy i popraw je własną (sensowną) metodą.

💻 ZADANIE 2: **PCA**

Ściągnij czystą, bezbłędną bazę danych z irysami:

- ze strony przedmiotu
- lub z Internetu (Google: iris dataset csv)
- lub pobierając z paczki sklearn datasets.

Następnie wykorzystując technikę PCA, chcemy skompresować bazę danych z czterech do trzech lub dwóch kolumn numerycznych, nie tracąc przy tym zbyt wiele informacji.

Wybierz paczkę i samouczek do zadania, np.:

- https://scikit-learn.org/stable/auto_examples/decomposition/plot_pca_iris.html
- https://notebook.community/hershaw/data-science-101/course/class1/pca/iris/PCA%20-%20Iris%20dataset
- https://builtin.com/machine-learning/pca-in-python

Możesz też skorzystać z programu pokazanego na wykładzie:

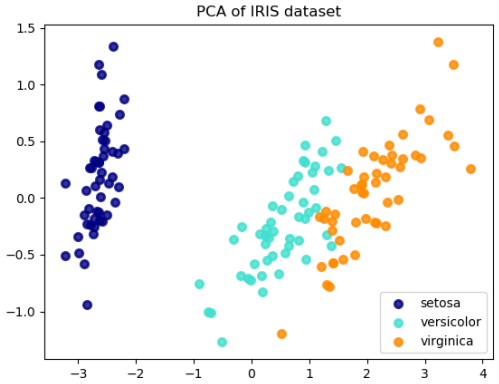
```
from sklearn import datasets
from sklearn.decomposition import PCA
import pandas as pd

iris = datasets.load_iris()
X = pd.DataFrame(iris.data, columns=iris.feature_names)
y = pd.Series(iris.target, name='FlowerType')
print(X.head())

pca_iris = PCA(n_components=3).fit(iris.data)
print(pca_iris)
print(pca_iris.explained_variance_ratio_)
print(pca_iris.components_)
print(pca_iris.transform(iris.data))
```

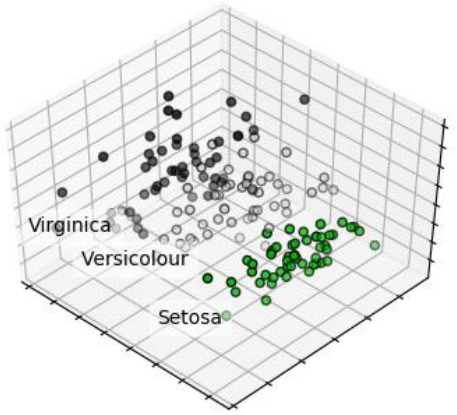
Dokonaj PCA na bazie danych. Przyjrzyj się nowym kolumnom i wariancjom. Ile kolumn można usunąć, tak aby zachować minimum 95% wariancji (strata informacji nie może być większa niż 5%)? Korzystając z poniższego wzoru, swoją odpowiedź uzasadnij.

Grzegorz Madejski
grzegorz.madejski@ug.edu.pl
Zakład Sztucznej Inteligencji,
Instytut Informatyki, Uniwersytet Gdański

$$Strata\ informacji\ spowodowana\ usunięciem\ i\ ostatnich\ \ kolumn = \frac{\sum_{k=n-i}^{n-1} Var(kolumna[k])}{\sum_{k=0}^{n-1} Var(kolumna[k])}$$

Bazę danych z usuniętymi kolumnami zobrazuj na wykresie punktowym, gdzie każdy punkt to irys. Jeśli w bazie zostawisz 2 kolumny, to wykres będzie na płaszczyźnie, a jeśli 3, to będzie trójwymiarowy. Przykłady:



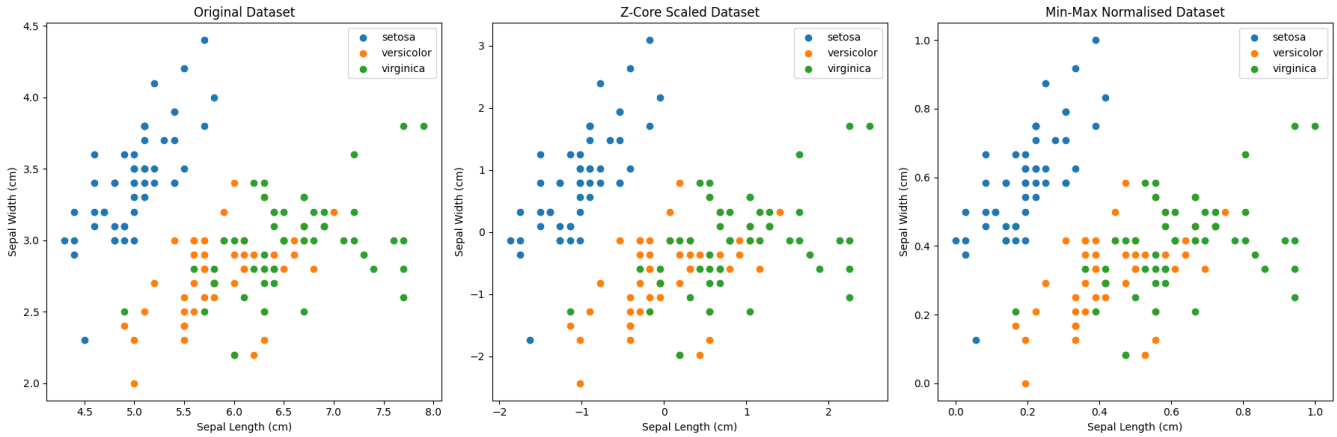


ZADANIE 3: **NORMALIZACJA**

Najpierw spróbuj wykonać zadanie samodzielnie, a następnie wykorzystaj ChatGPT, lub Google Gemini lub inne narzędzie do rozwiązania tego zadania i zapisz użyte przez Ciebie prompty. Zadanie:

Stwórz wykresy z irysami jako punktami na wykresie, dla dwóch zmiennych: sepal length i sepal width. Klasy irysów oznaczone są w legendzie wykresu. Zrób wykres w trzech wersjach: dane oryginalne, znormalizowane min-max i zeskalowane z-scorem. Wynik powinien przypominać ten poniżej.

Co możesz powiedzieć o min, max, mean, standard deviation dla tych danych?



Grzegorz Madejski
grzegorz.madejski@ug.edu.pl
Zakład Sztucznej Inteligencji,
Instytut Informatyki, Uniwersytet Gdański