WYŻSZA SZKOŁA
INFORMATYKI i ZARZĄDZANIA
z siedzibą w Rzeszowie

Rzeszów, 2023

ZARZĄDZANIE DANYMI

LABORATORIUM nr 5 (część 1)
Temat: Przetwarzanie danych wejściowych dla metod sztucznej inteligencji / uczenia maszynowego (moduł ***scikit-learn***)

Laboratorium obejmuje implementację skryptów w języku ***Python***, umożliwiających wstępne przetwarzanie danych wyjściowych, w celu ich dalszego wykorzystania w procesie uczenia klasyfikatorów ***AI/ML*** (ang., *Artificial Intelligence / Machine Learning*). Zadania obejmują kolejno:

- obsługę brakujących wartości,
- kodowanie wartości atrybutów (dwu- lub wielowartościowych),
- podział zbioru wejściowego na zbiór danych ***treningowych*** (uczących) i ***testowych***, oraz
- skalowanie wartości/cech,

przy pomocy biblioteki ***scikit-learn*** z poziomu środowiska ***Jupyter Notebook/LAB***.

Samodzielne wykonanie zadań z laboratorium będzie wymagane z zastosowaniem środowiska ***Jupyter Notebook*** oraz/lub ***Jupyter LAB*** (pliki ****.ipynb***).

1. Import bibliotek oraz wczytywania zbioru danych wejściowych

	Country	Age	Salary	Purchased
0	France	44.0	72000.0	No
1	Spain	27.0	48000.0	Yes
2	Germany	30.0	54000.0	No
3	Spain	38.0	61000.0	No
4	Germany	40.0	NaN	Yes
5	France	35.0	58000.0	Yes
6	Spain	NaN	52000.0	No
7	France	48.0	79000.0	Yes
8	Germany	50.0	83000.0	No
9	France	37.0	67000.0	Yes

Rys. 1. Źródłowy zbiór danych wejściowych (z ***brakującymi*** wartościami)

```
[['France' 44.0 72000.0]
 ['Spain' 27.0 48000.0]
 ['Germany' 30.0 54000.0]
 ['Spain' 38.0 61000.0]
 ['Germany' 40.0 nan]
 ['France' 35.0 58000.0]
 ['Spain' nan 52000.0]
 ['France' 48.0 79000.0]
 ['Germany' 50.0 83000.0]
 ['France' 37.0 67000.0]]
```

Rys. 2.A. Źródłowy zbiór danych wejściowych – ***atrybuty warunkowe*** (***X***)

```
['No' 'Yes' 'No' 'No' 'Yes' 'Yes' 'No' 'Yes' 'No' 'Yes']
```

Rys. 2.B. Źródłowy zbiór danych wejściowych – ***atrybut decyzyjny*** (***y***)

2. Obsługa brakujących wartości

```
[['France' 44.0 72000.0]
 ['Spain' 27.0 48000.0]
 ['Germany' 30.0 54000.0]
 ['Spain' 38.0 61000.0]
 ['Germany' 40.0 63777.77777777778]
 ['France' 35.0 58000.0]
 ['Spain' 38.77777777777778 52000.0]
 ['France' 48.0 79000.0]
 ['Germany' 50.0 83000.0]
 ['France' 37.0 67000.0]]
```

Rys. 3. Zbiór danych z uzupełnionymi (zastąpionymi) wartościami

3. Kodowanie (dekodowanie) danych

```
[[1.0 0.0 0.0 44.0 72000.0]
 [0.0 0.0 1.0 27.0 48000.0]
 [0.0 1.0 0.0 30.0 54000.0]
 [0.0 0.0 1.0 38.0 61000.0]
 [0.0 1.0 0.0 40.0 63777.77777777778]
 [1.0 0.0 0.0 35.0 58000.0]
 [0.0 0.0 1.0 38.77777777777778 52000.0]
 [1.0 0.0 0.0 48.0 79000.0]
 [0.0 1.0 0.0 50.0 83000.0]
 [1.0 0.0 0.0 37.0 67000.0]]
```

Rys. 4.A. Zamiana (kodowanie) wartości dla atrybutu z lokalizacjami (kolumna ***Country***)

```
print(y)
['No' 'Yes' 'No' 'No' 'Yes' 'Yes' 'No' 'Yes' 'No' 'Yes']

le = LabelEncoder()
y = le.fit_transform(y)

print(y)
[0 1 0 0 1 1 0 1 0 1]
```

Rys. 4.B. Zamiana (kodowanie) wartości dla atrybutu decyzyjnego (kolumna ***Purchased***)

4. Podział zbioru danych na zbiór treningowy oraz zbiór testowy

```
X_train
```

```
array([[0.0, 0.0, 1.0, 38.77777777777778, 52000.0],
       [0.0, 1.0, 0.0, 40.0, 63777.77777777778],
       [1.0, 0.0, 0.0, 44.0, 72000.0],
       [0.0, 0.0, 1.0, 38.0, 61000.0],
       [0.0, 0.0, 1.0, 27.0, 48000.0],
       [1.0, 0.0, 0.0, 48.0, 79000.0],
       [0.0, 1.0, 0.0, 50.0, 83000.0],
       [1.0, 0.0, 0.0, 35.0, 58000.0]], dtype=object)
```

```
X_test
```

```
array([[0.0, 1.0, 0.0, 30.0, 54000.0],
       [1.0, 0.0, 0.0, 37.0, 67000.0]], dtype=object)
```

Rys. 5.A. Podział wejściowego zbioru danych na część ***treningową*** i ***testową*** (kolumny dla ***atrybutów warunkowych***)

```
y_train
```

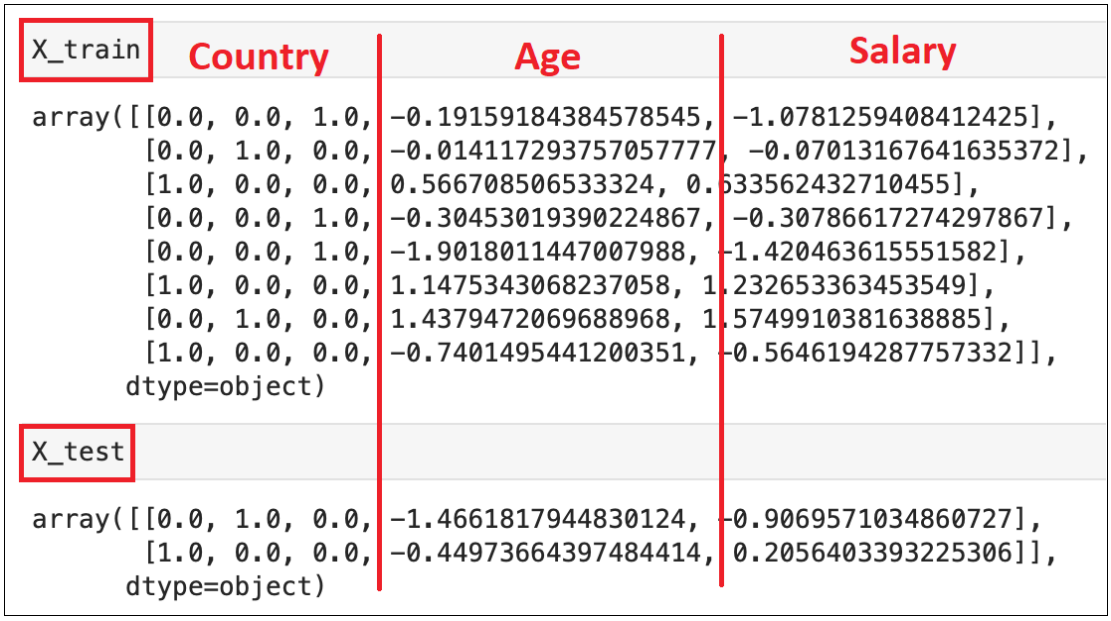
```
array([0, 1, 0, 0, 1, 1, 0, 1])
```

```
y_test
```

```
array([0, 1])
```

Rys. 5.B. Podział wejściowego zbioru danych na część ***treningową*** i ***testową*** (kolumna z ***atrybutem decyzyjnym***)

5. Skalowanie cech (wartości)



Rys. 5.B. Przykład kolumn z ***atrybutami decyzyjnymi***) po procesie skalowania ***wartości / cech***