Akademia Ekonomiczno-Humanistyczna w Warszawie

SPRAWOZDANIE

INTELIGENTNA ANALIZA DANYCH

**LAB2**

**WALIDACJA KRZYŻOWA, STRATYFIKACJA,**

**DRZEWA LOSOWE I NAIWNY KLASYFIKATOR bayesowski**

07.11.2021

JOANNA PRAJZENDANC

36358

MIŁOSZ SAKOWSKI

36381

**Spis treści**

[1. Cel i przebieg ćwiczenia 3](#_Toc29772)

[2. Definicje i założenia 3](#_Toc13946)

[2.1. Wyjaśnienie pojęć 3](#_Toc23313)

[2.2. Przygotowanie bazy danych 3](#_Toc28633)

[i. Importowanie bazy danych i oryginalna zawartość 3](#_Toc7684)

[ii. Docelowa baza danych 3](#_Toc12652)

[3. Walidacja krzyżowa i stratyfikacja 3](#_Toc13095)

[3.1. Zadanie #1 3](#_Toc30990)

[i. Treść polecenia 3](#_Toc2373)

[ii. Rozwiązanie 3](#_Toc10574)

[4. Klasyfikator drzewa losowego 4](#_Toc13270)

[4.1. Zadanie #2 4](#_Toc32327)

[i. Omówienie kodu 4](#_Toc8963)

[ii. Pytania 4](#_Toc8720)

[4.2. Zadanie #3 4](#_Toc12503)

[i. Treść polecenia 4](#_Toc9646)

[ii. Rozwiązanie 4](#_Toc21720)

[4.3. Wnioski 4](#_Toc2746)

[5. Naiwny klasyfikator bayesowski 4](#_Toc6567)

[5.1. Zadanie #4 4](#_Toc11031)

[i. Treść polecenia 4](#_Toc3155)

[ii. Rozwiązanie 5](#_Toc30568)

[5.2. Wnioski 5](#_Toc23423)

[6. Porównanie klasyfikatorów 5](#_Toc2600)

[7. Wnioski końcowe 5](#_Toc19602)

# Cel i przebieg ćwiczenia

Celem ćwiczenia było utrwalenie wiedzy w zakresie klasyfikatorów. W pierwszej kolejności należało zapoznać się z problemem niezbalansowanego zbioru danych i sposobów rozwiązania go za pomocą walidacji krzyżowej i stratyfikacji w przypadku klasyfikatora binarnego. Następnym krokiem było zapoznanie się z klasyfikatorem drzewa losowego oraz naiwnym klasyfikatorem bayesowski.

# Definicje i założenia

## Wyjaśnienie pojęć

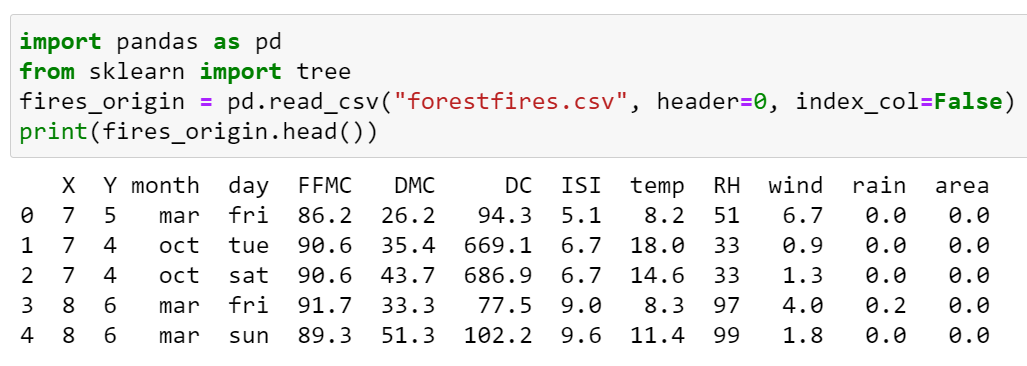
W sprawozdaniu pojawiają się następujące pojęcia:

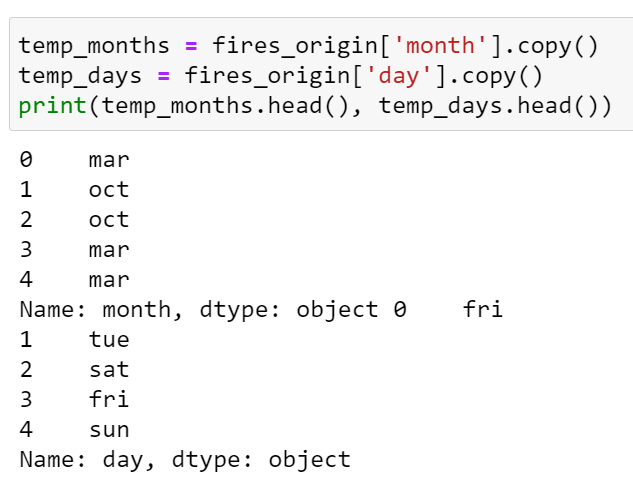
* niezbalansowana baza danych: zbiór danych, w którym występuje wyraźna dysproporcja jeśli chodzi o ilość elementów dla każdej z klas,
* walidacja krzyżowa:
* stratyfikacja
* klasyfikator binarny
* klasyfikator drzewa losowego
* naiwny klasyfikator bayesowski

## Przygotowanie bazy danych

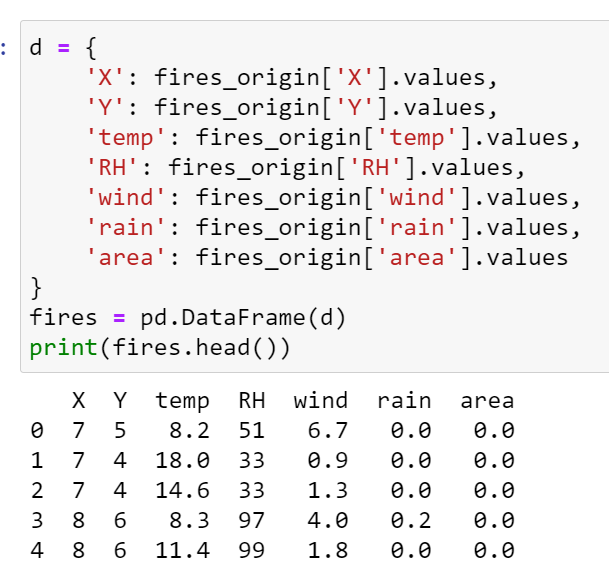
Podczas wykonywania zadań omówionych w sprawozdaniu, do rozwiązania użyto tej samej bazy danych z informacjami o pożarach w parku XX.

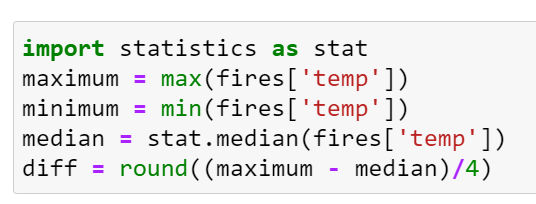
### Importowanie bazy danych i oryginalna zawartość

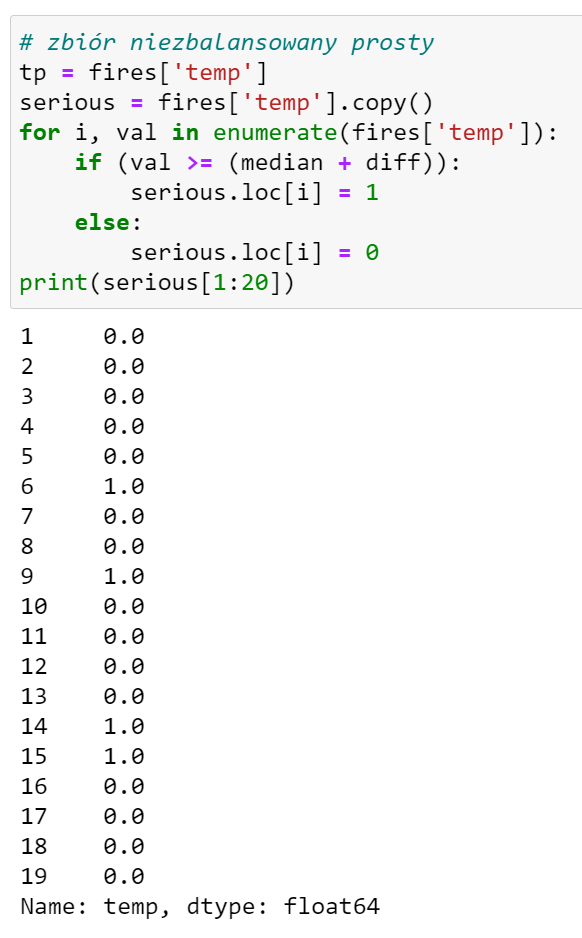


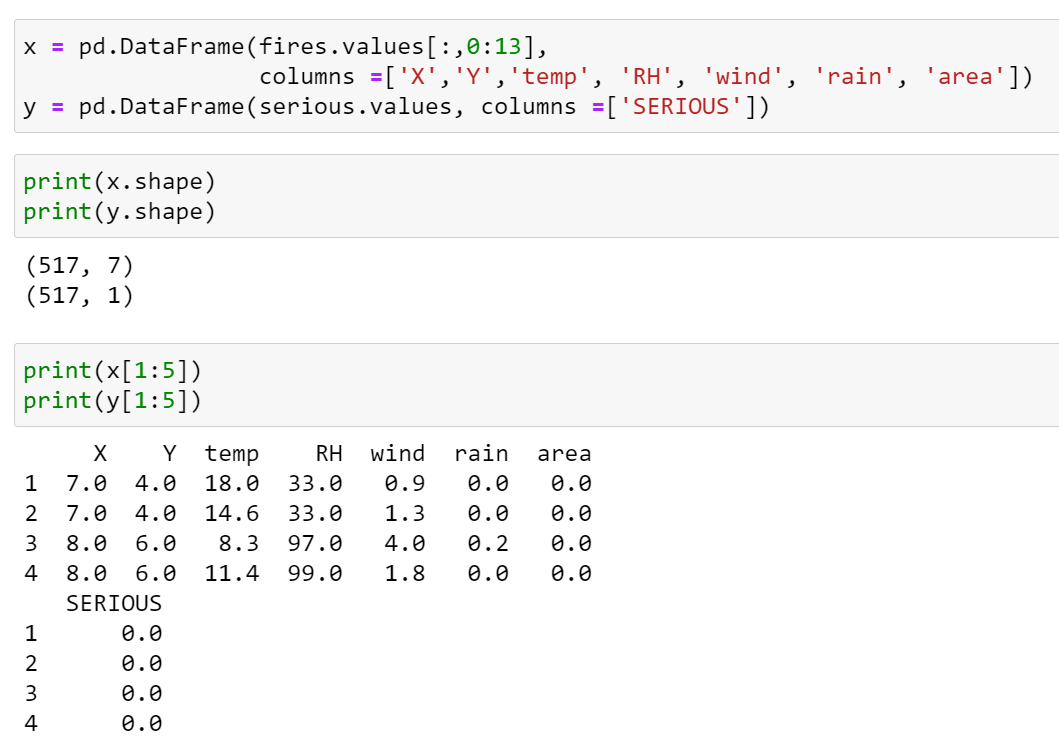


### Docelowa baza danych









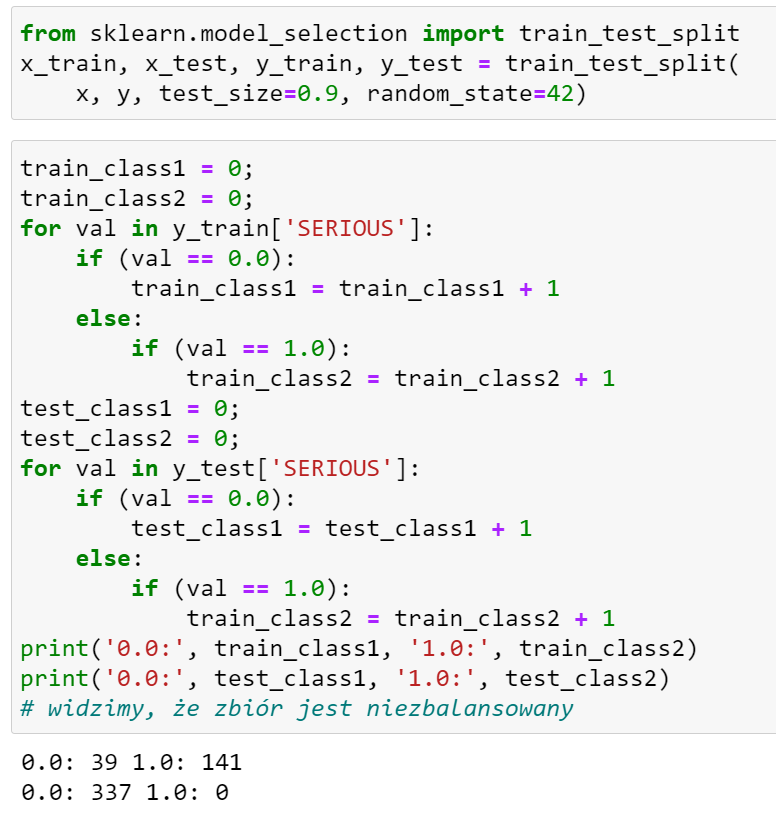
# Walidacja krzyżowa i stratyfikacja

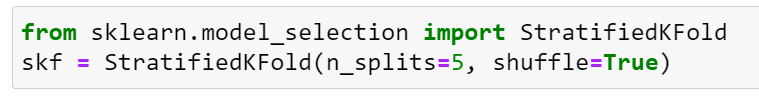
## Zadanie #1

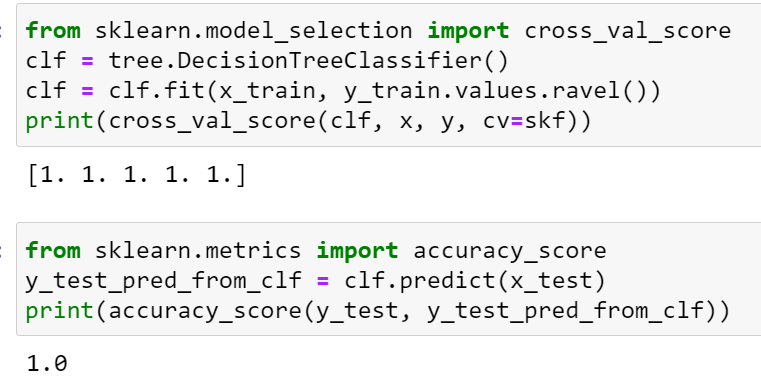
### Treść polecenia

Proszę pobrać inny zbiór danych, zmienić klasy w taki sposób żeby zbiór był niezbalansowany oraz przetestować drzewo decyzyjne stosując 5-krotną cross walidację ze stratyfikacją

### Rozwiązanie



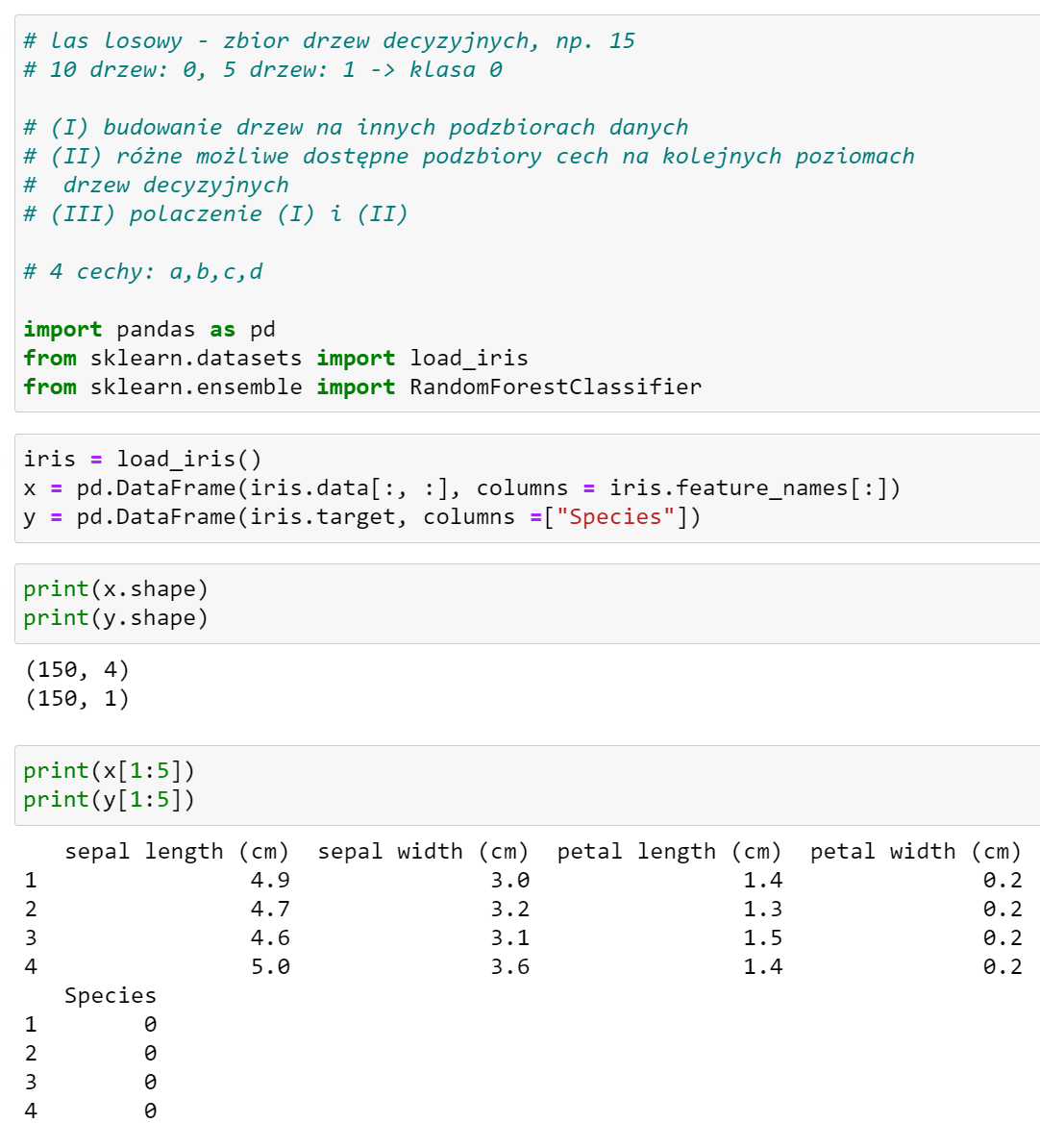


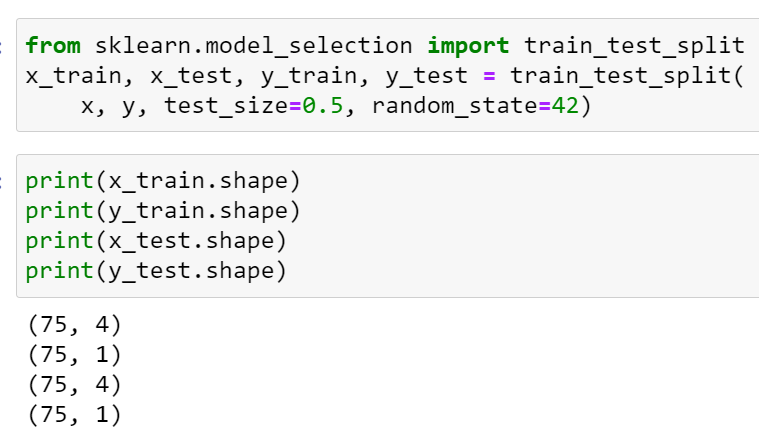


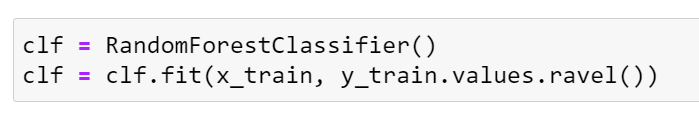
# Klasyfikator drzewa losowego

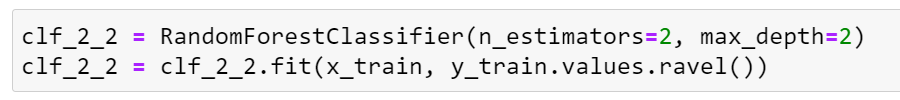
## Zadanie #2

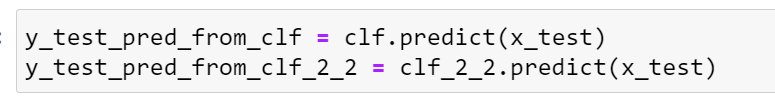
### Omówienie kodu

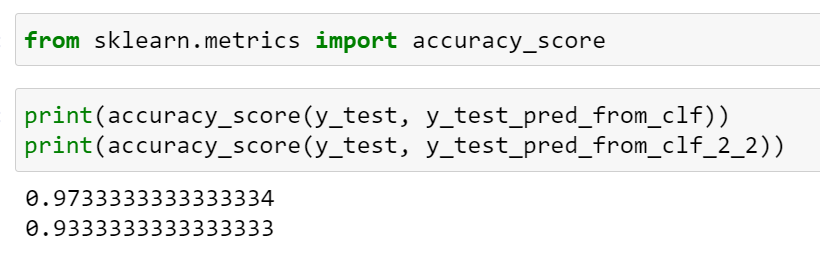












### Pytania

#### Czy uzyskane wyniki są zadowalające?

Na pierwszy rzut oka wszystko wygląda w porządku i dokładność jest na wystarczającym poziomie (o ile nie potrzebujemy bardzo wysokiej precyzji wyniku).

#### Czy model został przetestowany w prawidłowy sposób?

Nie, ponieważ istnieje ryzyko, że w zbiorze trenującym są same „0” i żadnych „1” albo w zbiorze testującym nie ma żadnych „1” więc nie zbadamy skuteczności modelu w dopasowywaniu „1”.

## Zadanie #3

### Treść polecenia

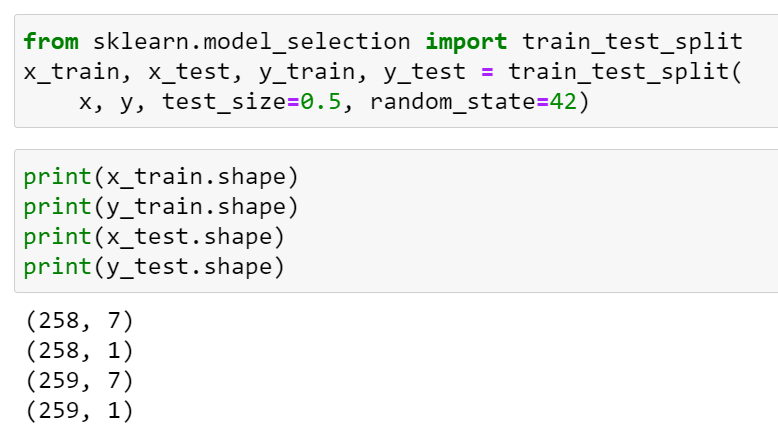
Proszę pobrać dowolny zbiór danych ze strony <https://archive.ics.uci.edu/ml/index.php.> Następnie proszę podzielić zbiór na dane trenujące i testujące, wytrenować 5 modeli lasów losowych z różną maksymalną głębokością i liczbą drzew w lesie, porównać wyniki. Proszę o sporządzenie sprawozdania z wnioskami.

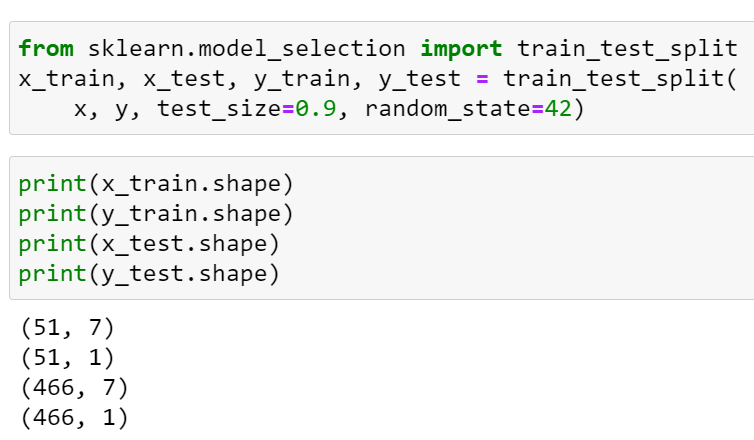
### Rozwiązanie

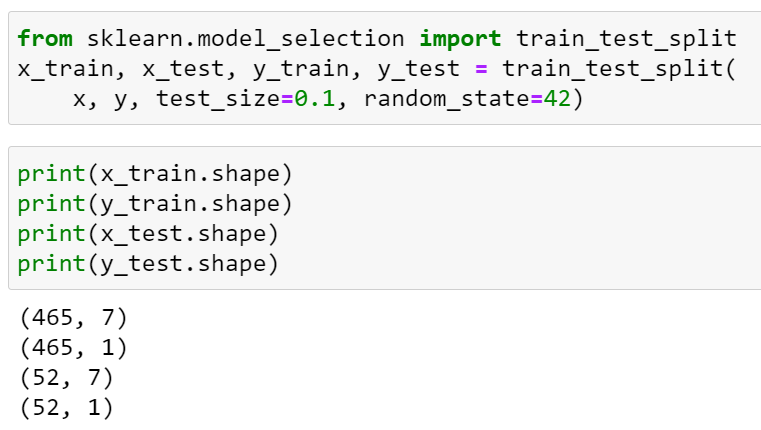
#### Adnotacja odnośnie podziału na zbiór testujący i trenujący

Podczas wykonywania zadań, zauważono że proporcja zbioru testującego do trenującego podczas dokonywania podziału ma duże znaczenie, dlatego podczas zbierania wyników uwzględniono trzy różne sposoby podzielenia:

* testujący 10% i trenujący 90%: test\_size = 0.1
* testujący 50% i trenujący 50%: test\_size = 0.5
* testujący 90% i trenujący 10%: test\_size = 0.9



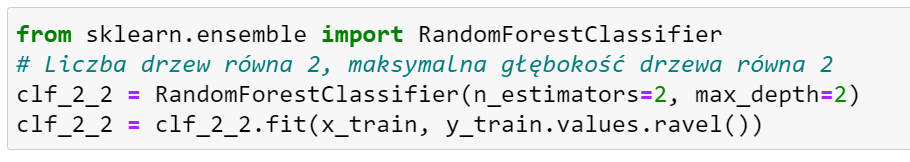


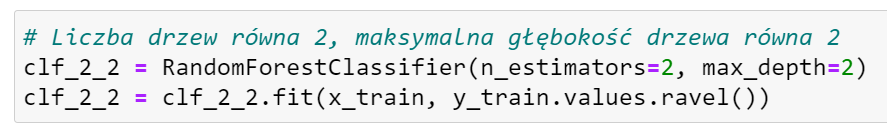


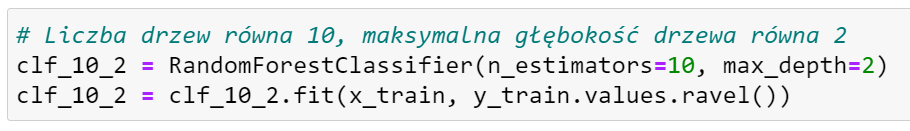
#### Budowanie lasów losowych

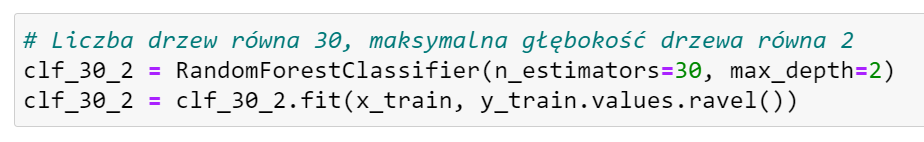
Zbudowane lasy:

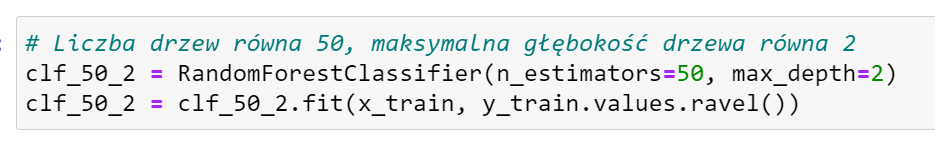
* 2 drzewa o maksymalnej głębokości równej 2,
* 10 drzew o maksymalnej głębokości równej 2,
* 30 drzew o maksymalnej głębokości równej 2,
* 50 drzew o maksymalnej głębokości równej 2,
* 70 drzew o maksymalnej głębokości równej 2
* 10 drzew o maksymalnej głębokości równej 1,
* 10 drzew o maksymalnej głębokości równej 3.

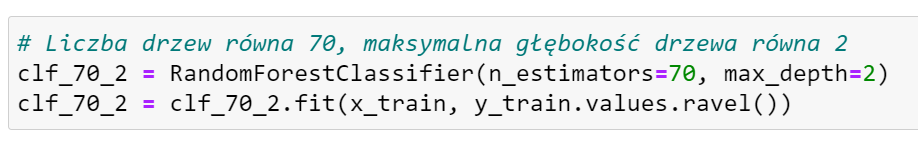


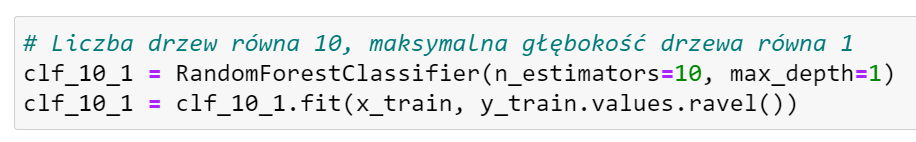


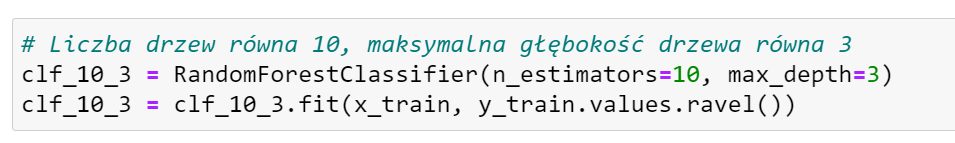


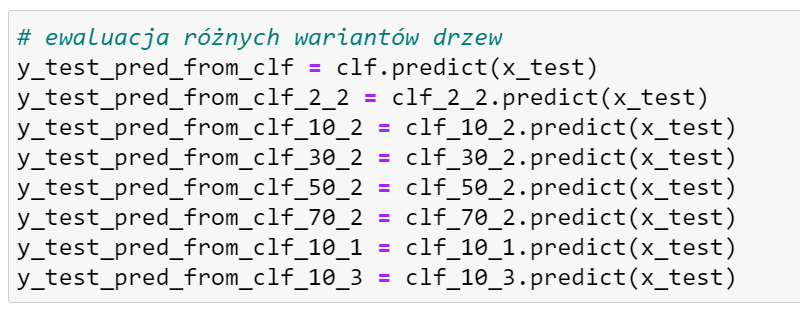




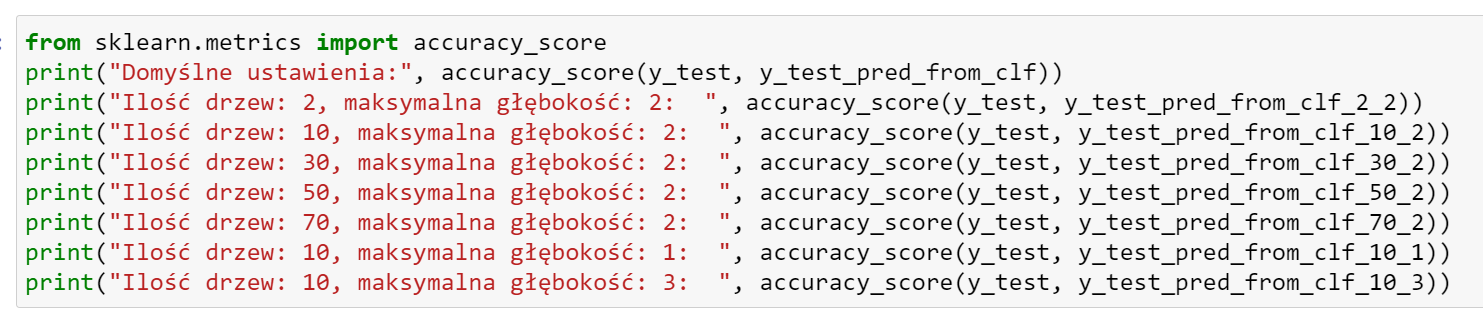




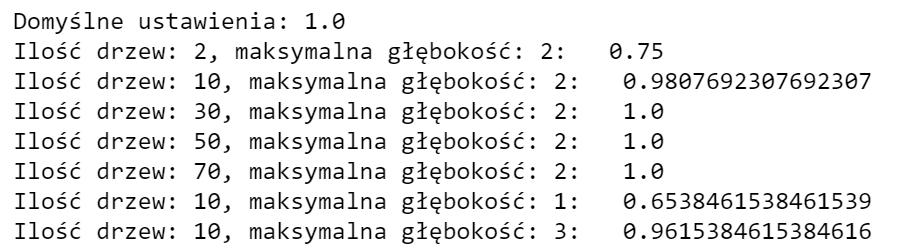




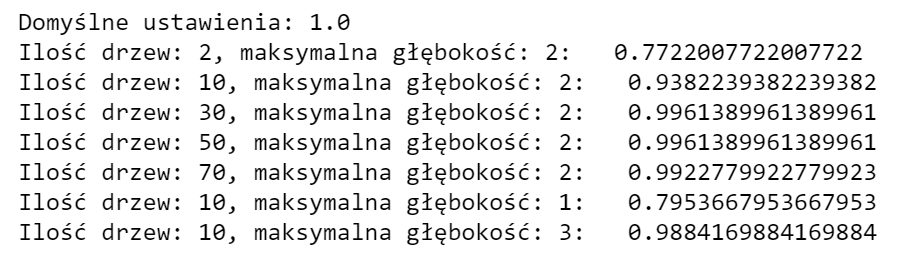
#### Porównanie wyników



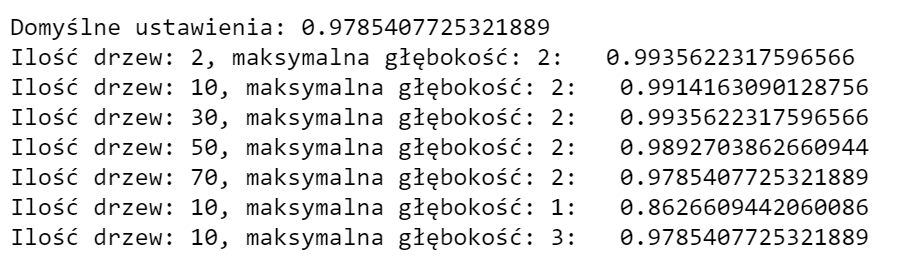
Test\_size: 0.1



Test\_size: 0.5



Test\_size: 0.9



Maksymalna głębokość: 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test\_size = 0.1 | Test\_size = 0.5 | Test\_size = 0.9 |
| Domyślnie (100 drzew i nieograniczona głębokość) | 1.0 | 1.00 | 0.98 |
| 10 drzew | 0.98 | 0.938 | 0.99 |
| 30 drzew | 1.0 | 0.996 | 0.99 |
| 50 drzew | 1.0 | 0.996 | 0.99 |
| 70 drzew | 1.0 | 0.992 | 0.98 |

Ilość drzew: 10

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test\_size = 0.1 | Test\_size = 0.5 | Test\_size = 0.9 |
| Domyślnie (100 drzew i nieograniczona głębokość) | 1.0 | 1.00 | 0.98 |
| Głębokość = 1 | 0.65 | 0.80 | 0.86 |
| Głębokość = 2 | 0.98 | 0.94 | 0.99 |
| Głębokość = 3 | 0.96 | 0.99 | 0.98 |

## Wnioski

# Naiwny klasyfikator bayesowski

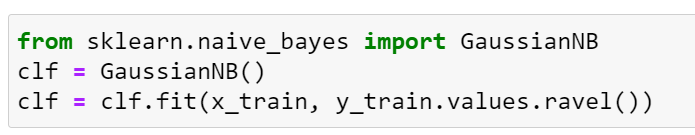
## Zadanie #4

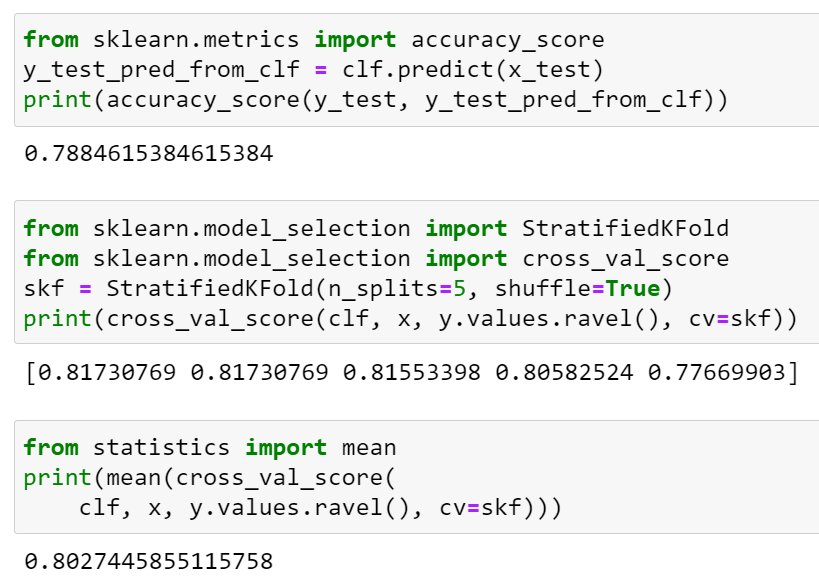
### Treść polecenia

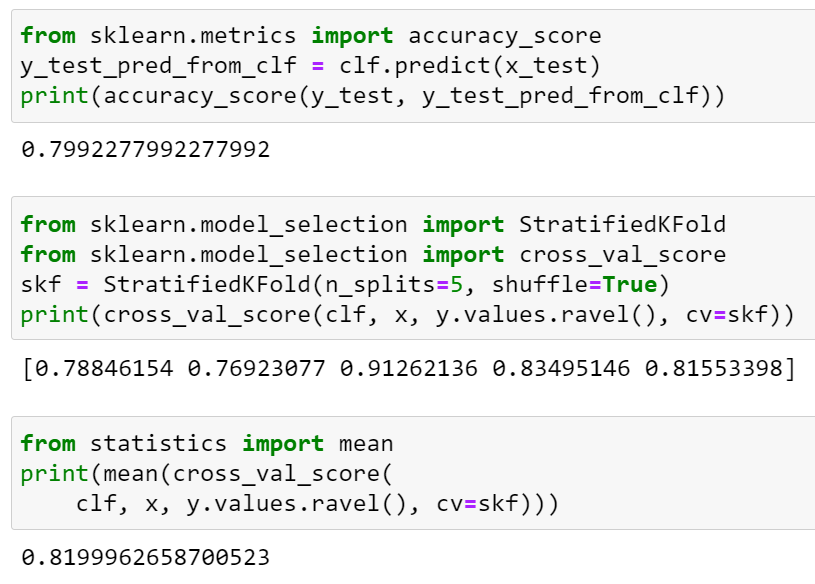
Proszę pobrać dowolny zbiór danych ze strony <https://archive.ics.uci.edu/ml/index.php> Następnie proszę podzielić zbiór na dane trenujące i testujące, wytrenować i przetestować klasyfikator GaussianNB. Proszę o sporządzenie sprawozdania z wnioskami.

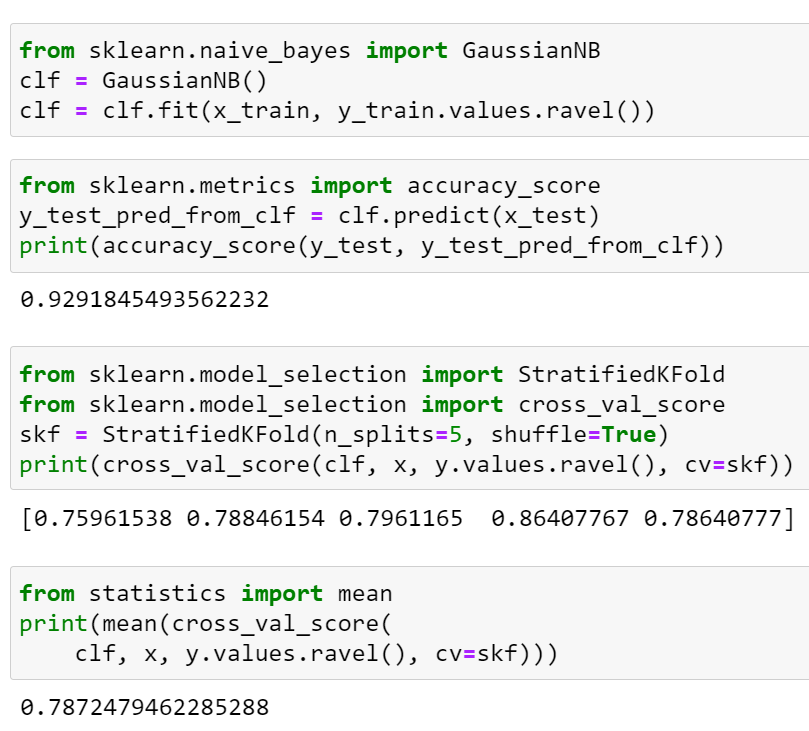
### Rozwiązanie

Podczas dokonywania obliczeń dla naiwnego klasyfikatora bayesowski również przetestowano 3 podziały na zbiór testujący i trenujący, jak w przypadku pomiarów dla [Zadanie #3](#_Zadanie #3).









|  |  |  |  |
| --- | --- | --- | --- |
|  | Test\_size = 0.1 | Test\_size = 0.5 | Test\_size = 0.9 |
| Naiwny klasyfikator bayesowski | 0.79 | 0.80 | 0.93 |

## Wnioski

# Porównanie klasyfikatorów

|  |  |  |  |
| --- | --- | --- | --- |
|  | Test\_size = 0.1 | Test\_size = 0.5 | Test\_size = 0.9 |
| Klasyfikator lasu losowego (100 drzew i nieograniczona głębokość) | 1.0 | 1.00 | 0.98 |
| Naiwny klasyfikator bayesowski | 0.79 | 0.80 | 0.93 |

# Wnioski końcowe