Akademia Ekonomiczno-Humanistyczna w Warszawie

SPRAWOZDANIE

INTELIGENTNA ANALIZA DANYCH

**LAB4**

**SIECI NEURONOWE I PORÓWNANIE KLASYFIKATORÓW**

12.12.2021

JOANNA PRAJZENDANC

36358

MIŁOSZ SAKOWSKI

36381

**Spis treści**

[1. Cel i przebieg ćwiczenia 3](#_Toc27127)

[2. Definicje i założenia 3](#_Toc6504)

[2.1. Wyjaśnienie pojęć 3](#_Toc12478)

[3. Sieci neuronowe 3](#_Toc5292)

[3.1. Omówienie klasyfikatora 3](#_Toc23101)

[3.2. Zadanie #1 4](#_Toc10219)

[i. Treść polecenia 4](#_Toc19562)

[ii. Rozwiązanie 4](#_Toc15427)

[iii. Porównanie wyników i wnioski 7](#_Toc743)

[4. Porównanie klasyfikatorów 8](#_Toc8673)

[4.1. Zadanie #2 8](#_Toc3061)

[i. Treść polecenia 8](#_Toc22943)

[ii. Baza danych 8](#_Toc27168)

[iii. Klasyfikator: binarny 12](#_Toc13456)

[iv. Klasyfikator: drzewo decyzyjne 13](#_Toc26650)

[v. Klasyfikator: las losowy 14](#_Toc16605)

[vi. Klasyfikator: naiwny bayesowski 15](#_Toc6303)

[vii. Klasyfikator: sieci neuronowe 16](#_Toc30374)

[4.2. Podsumowanie i wnioski 17](#_Toc19039)

# Cel i przebieg ćwiczenia

Celem ćwiczenia było utrwalenie wiedzy w zakresie poznanych dotychczas klasyfikatorów: klasyfikator binarny, klasyfikator drzewa decyzyjnego, naiwny klasyfikator bayesowski, klasyfikator lasu losowego i klasyfikator sieci neuronowych. Podczas wykonywania zadań zastosowano również poznaną wcześniej stratyfikację, walidację krzyżową i optymalizację modelu grid search.

# Definicje i założenia

## Wyjaśnienie pojęć

W sprawozdaniu pojawiają się nowe pojęcia:

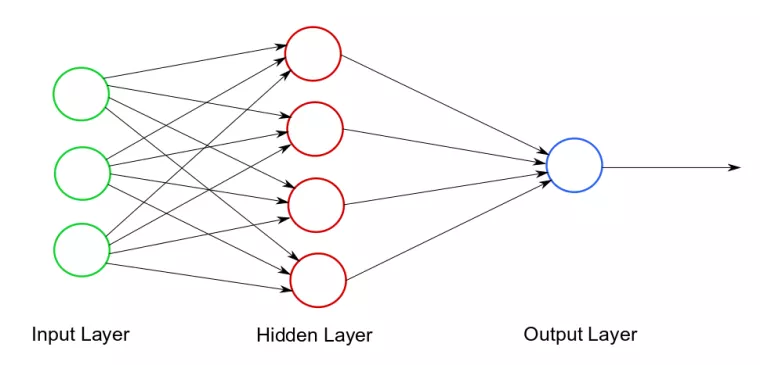
* sieci neuronowe - statystyczny model obliczeniowy stosowany w uczeniu maszynowym używany jako jeden z klasyfikatorów,
* neuron - warstwy sieci neuronowej, wyróżniamy 3 warstwy: warstwy wejścia (input layer), warstwy ukryte (hidden layer), oraz warstwy wyjścia (output layer).

# Sieci neuronowe

## Omówienie klasyfikatora

Sieć neuronowa to statystyczny model obliczeniowy stosowany w uczeniu maszynowym. Można o nim myśleć jak o systemie połączonych synapsami neuronów, które przesyłają między sobą impulsy (dane). Sieć neuronowa składa się z trzech warstw:

* warstwy wejścia (input layer),
* warstwy ukrytej (hidden layer),
* oraz warstwy wyjścia (output layer).



Obraz 1: Rysunek schematyczny sieci neuronowej z wyróżnionymi warstwami

Warstwa wejścia przyjmuje dane wejściowe do obliczeń, w warstwie ukrytej odbywają się wszystkie obliczenia. Wynik tych obliczeń jest przesyłany do warstwy wyjścia.

Na powyższym diagramie okręgi reprezentują neurony, zaś strzałki - synapsy. Każda synapsa ma przypisaną pewną wagę, tzn. liczbę, która (nieco upraszczając) określa, jak silnie przesyłana wartość wpływa na ostateczny wynik obliczeń. Żeby przesłać wartość, synapsa najpierw czyta wartość z neuronu wejściowego, następnie wartość tę mnoży przez wagę, by w końcu przesłać wynik do neuronu wyjściowego. Następnie neuron wyjściowy dokonuje obliczeń na dostarczonych mu przez synapsy wartościach i otrzymany wynik przekazuje do wychodzącej z niego synapsy.

## Zadanie #1

### Treść polecenia

Proszę pobrać dowolny zbiór danych ze strony <https://archive.ics.uci.edu/ml/index.php>

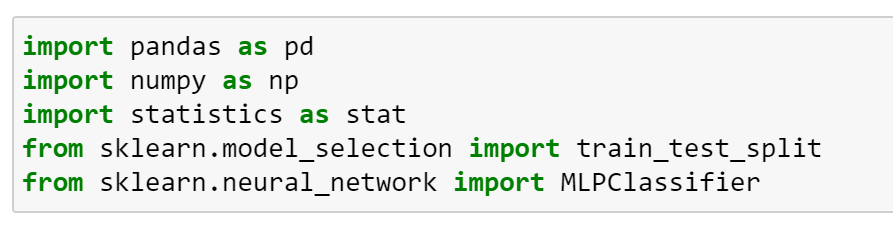
Następnie proszę podzielić zbiór na dane trenujące i testujące, wytrenować i przetestować 5 sieci neuronowych o różnych architekturach. Proszę o sporządzenie sprawozdania z wnioskami.

### Rozwiązanie

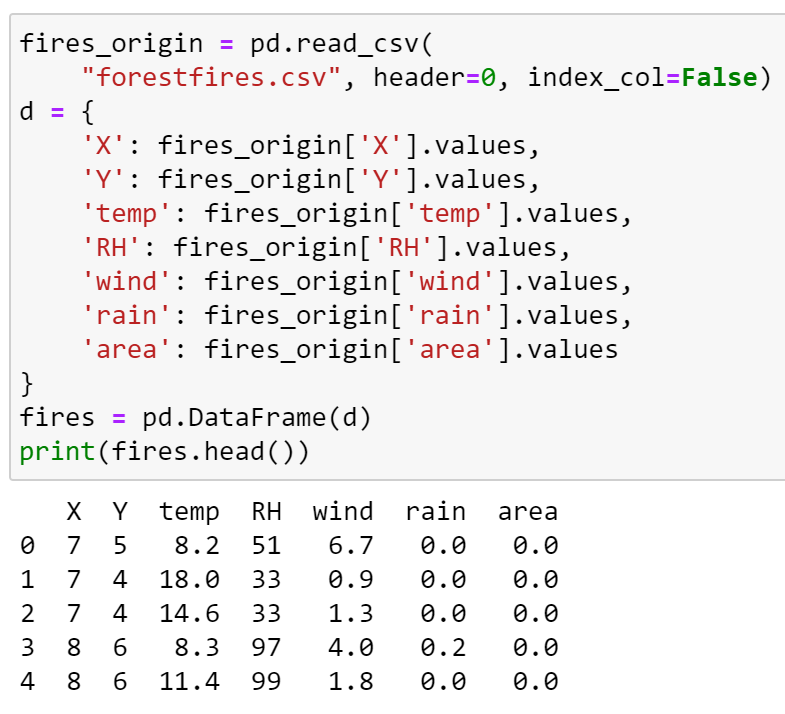
#### Przygotowanie danych

Dane wykorzystane w zadaniach pochodzą z pliku forestfires.csv, który został pobrany ze strony <https://archive.ics.uci.edu/ml/index.php>. Jest to zbiór informacji na temat pożarów w parku Montensinho.

Na potrzeby zadania pominięto kolumny z wartościami tekstowymi, ponieważ ich zamiana na wartości liczbowe była czasochłonna, a przydatność tych danych niewielka.



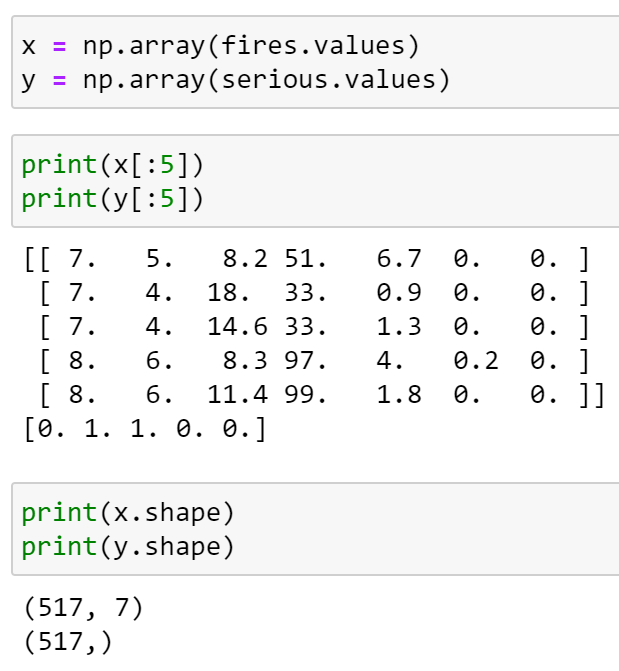
Obraz 2: Użyte biblioteki



Obraz 3: Wybrane kolumny z pliku forestfires.csv



Obraz 4: Utworzenie kolumny z danymi do klasyfikacji na podstawie własnego algorytmu



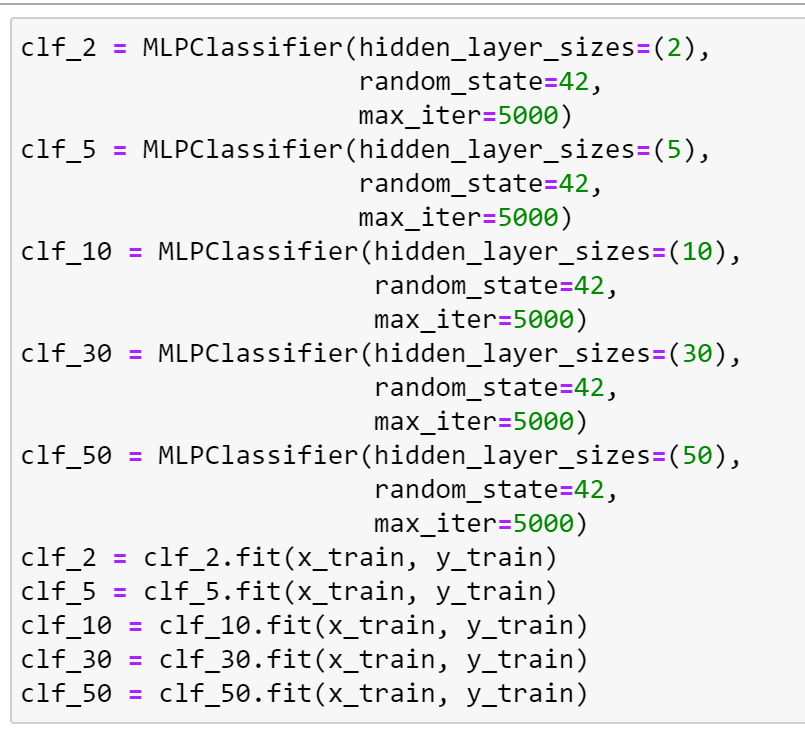
Obraz 5: Podział danych na zbiór trenujący i testujący

#### Wyniki dokładności sieci neuronowych z różną ilością neuronów

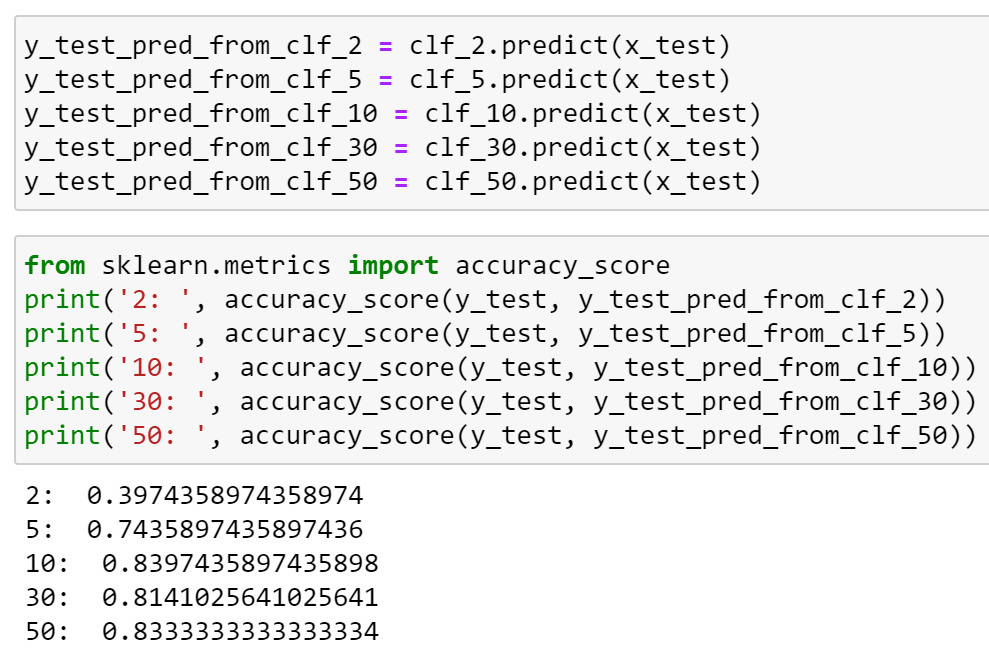
* hidden\_layer\_sizes - parametr ten reprezentuje liczbę neuronów w warstwie ukrytej.
* random\_state - określa generowanie liczb losowych na potrzeby inicjalizacji wag i odchyleń.
* max\_iter - parametr ten określa liczbę epok, ile razy każdy punkt danych zostanie użyty.

W zadaniu zbudowano 5 różnych sieci neuronowych:

1. Ilość neuronów: 2,
2. Ilość neuronów: 5,
3. Ilość neuronów: 10,
4. Ilość neuronów: 30,
5. Ilość neuronów: 50.



Obraz 6: Zastosowanie klasyfikatora sieci neuronowych z różną ilością neuronów



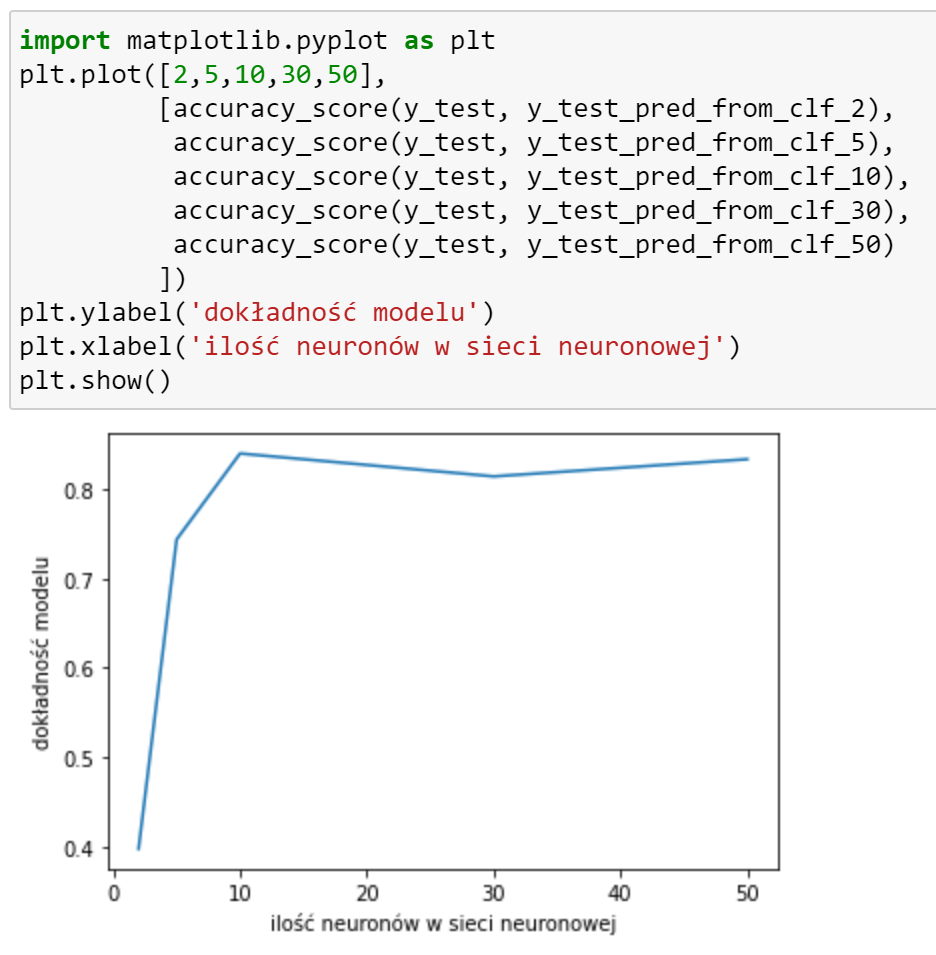
Obraz 7: Obliczenie dokładności modelu w zależności od ilości neuronów

### Porównanie wyników i wnioski

Tabela 1: Zestawienie wyników dokładności modelu sieci neuronowej w zależności od liczby neuronów

|  |  |
| --- | --- |
| **Ilość neuronów** | **Dokładność** |
| 2 | 0,397 |
| 5 | 0,744 |
| 10 | 0,840 |
| 30 | 0,814 |
| 50 | 0,833 |

Dokładność modelu wyraźnie rośnie w zależności od liczby neuronów w sieci neuronowej.



Obraz 8: Wykres zależności dokładności modelu od ilości neuronów

Prawdopodobnie zależność ta jest funkcją logarytmiczną, co oznacza że wartość dokładności szybko rośnie już dla niewielkiej liczby neuronów - dzięki czemu szybko możemy otrzymać zadowalające przybliżenie, ale też uzyskanie bardzo dużej dokładności (powyżej 90%) będzie wymagało sieci neuronów o bardzo dużej ilości neuronów, co przekłada się na czas obliczeń i duże ryzyko przetrenowania modelu.

# Porównanie klasyfikatorów

## Zadanie #2

### Treść polecenia

Pobrać wybraną bazę danych i porównać wytrenowanie różnymi klasyfikatorami.

### Baza danych

#### Informacje techniczne

* data dodania bazy danych: 22.07.2019,
* charakterystyka: wielowymiarowa,
* kategoria: komputery,
* ilość danych: 125,
* brakujące wartości: brak.

Bazę danych pobrano z <https://archive.ics.uci.edu/ml/datasets/Alcohol+QCM+Sensor+Dataset>

W pobranym zestawie danych znajdowało się 5 plików z wynikami pomiarów:

* QCM3.csv,
* QCM6.csv,
* QCM7.csv,
* QCM10.csv,
* QCM12.csv.

Nazwy plików odpowiadają numerowi czujnika QCM użytego podczas pomiarów.

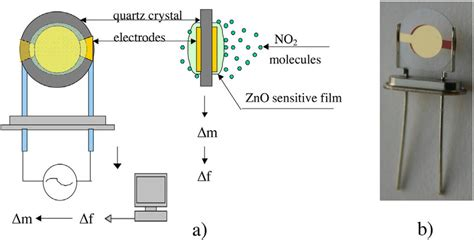
#### Opis badania i zawartość zestawu danych[[1]](#footnote-0)

Celem badania było zmierzenie reakcji różnych czujników QCM na 5 wybranych alkoholi, aby określić który z tych czujników będzie najlepszy do klasyfikacji tych alkoholi.

Pomiary dotyczyły 5 różnych gazów (alkoholi):

* 1-octanolu,
* 1-propanolu,
* 2-butanolu,
* 2-propanolu,
* 1-isobutanolu.

Wymienione gazy zostały zbadane przez 5 różnych sensorów QCM[[2]](#footnote-1), czyli przez mikrowagi kwarcowe, która są rodzajem czujnika do wykrywania bardzo małych zmian masy. Mikrowaga kwarcowa działa na zasadzie rezonatora kwarcowego pracującego z drganiami ścinającymi[[3]](#footnote-2) i jest używana do budowania tzw. *Elektronicznego nosa[[4]](#footnote-3)*.



Obraz 9: a) rysunek schematyczny mikrowagi kwarcowej; b) zdjęcie przykładowej mikrowagi kwarcowej

Rezonator kwarcowy składa się z dwóch okręgów, które różnią się zawartością MIP[[5]](#footnote-4) i NP[[6]](#footnote-5). Każdy z okręgów to osobny kanał pomiaru drgań (kanał pomiaru).

Tabela 2: Stosunki MIP i NP w każdym z czujników

|  |  |  |
| --- | --- | --- |
| **Czujnik** | **MIP** | **NP** |
| QCM3 | 1 | 1 |
| QCM6 | 1 | 0 |
| QCM7 | 1 | 0,5 |
| QCM10 | 1 | 2 |
| QCM12 | 0 | 1 |

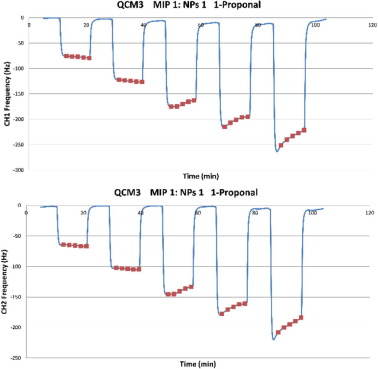
Podłączając mikrowagę do układu elektronicznego, można zmierzyć zmianę częstotliwości drgań rezonatora, która odpowiada zmianie masy. Dzięki temu można „zważyć” gaz.

Jeden pomiar dla jednego czujnika trwał 120 min, w trakcie tego czasu najpierw czujnik był umieszczany na 30 min w czystym powietrzu w celu oczyszczenia, następnie wybrany gaz był dodawany do powietrza aż do uzyskania zadanej koncentracji powietrze - alkohol i dokonywano pomiaru. Przed kolejnym pomiarem w innej koncentracji tego samego gazu czujnik był oczyszczany przez 7 min w czystym powietrzu.

Tabela 3: Wartości koncentracji powietrze - alkohol, dla których wykonano pomiary

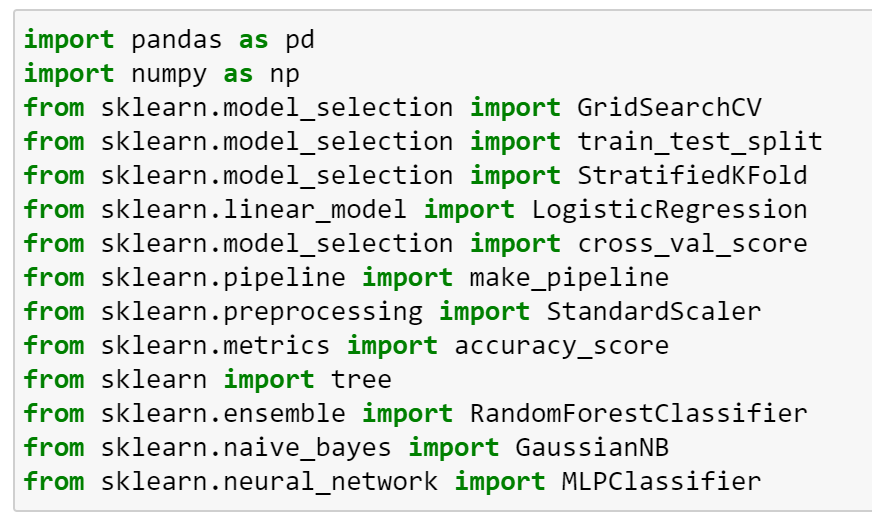
|  |  |  |
| --- | --- | --- |
| **L.p.** | **Koncentracja powietrza** | **Koncentracja alkoholu** |
| 1 | 0,799 | 0,201 |
| 2 | 0,700 | 0,300 |
| 3 | 0,600 | 0,400 |
| 4 | 0,501 | 0,499 |
| 5 | 0,400 | 0,600 |

Wszystkie pomiary zostały przeprowadzone w temperaturze pokojowej 25°C. Alkohol w stanie płynnym był przelewany do szklanej tuby o pojemności 50ml z umieszczonym czujnikiem. Próbka alkoholu docierała do czujnika jako gaz i wyniki zmiany częstotliwości drgań w Hz z każdego z dwóch kanałów mikrowagi kwarcowej były przesyłane do komputera.

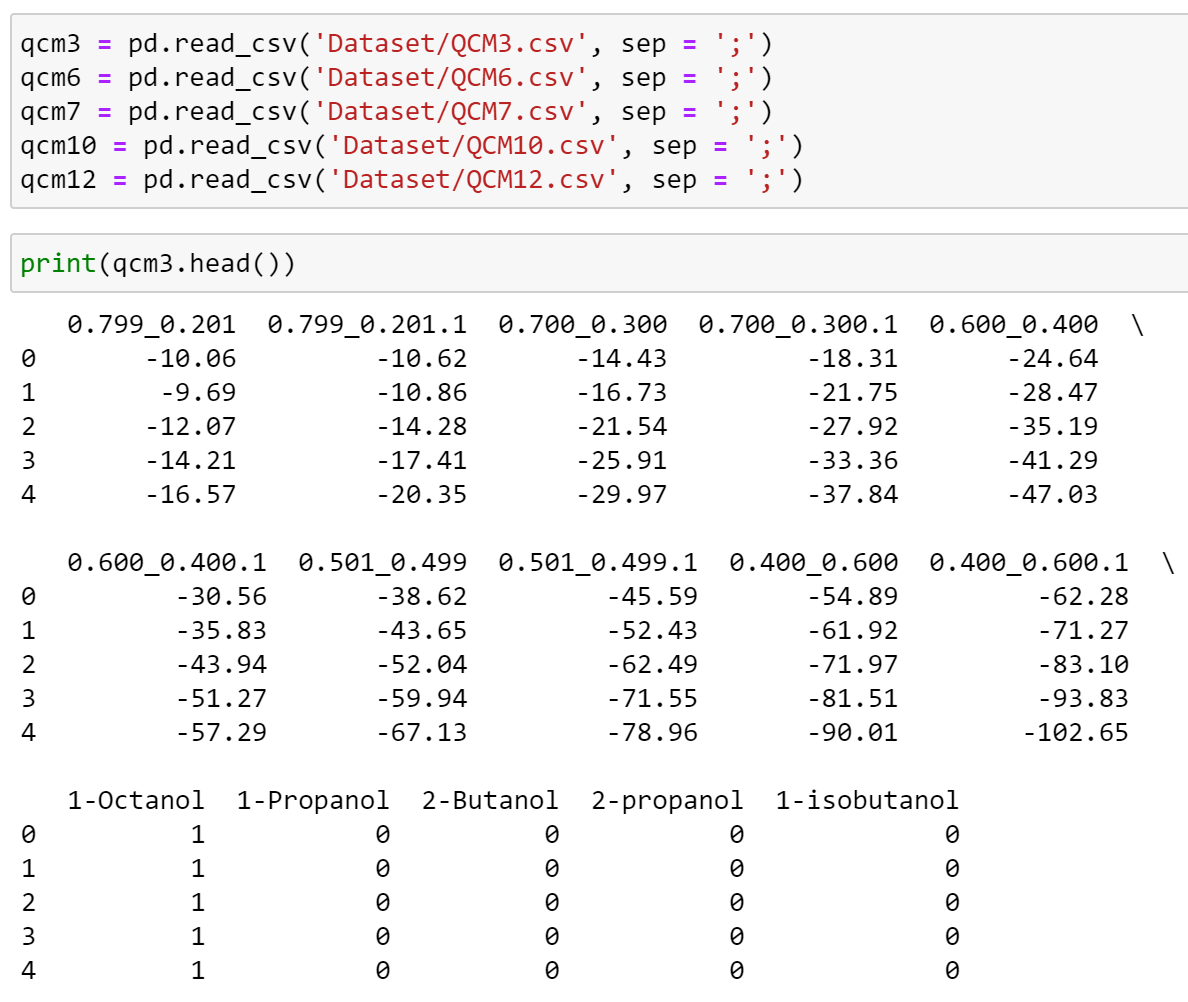


Obraz 10: Przykład pomiarów zmiany częstotliwości dla 1-Proponalu z czujnika QCM3

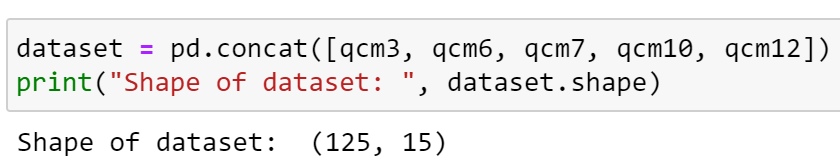
#### Przygotowanie do klasyfikacji



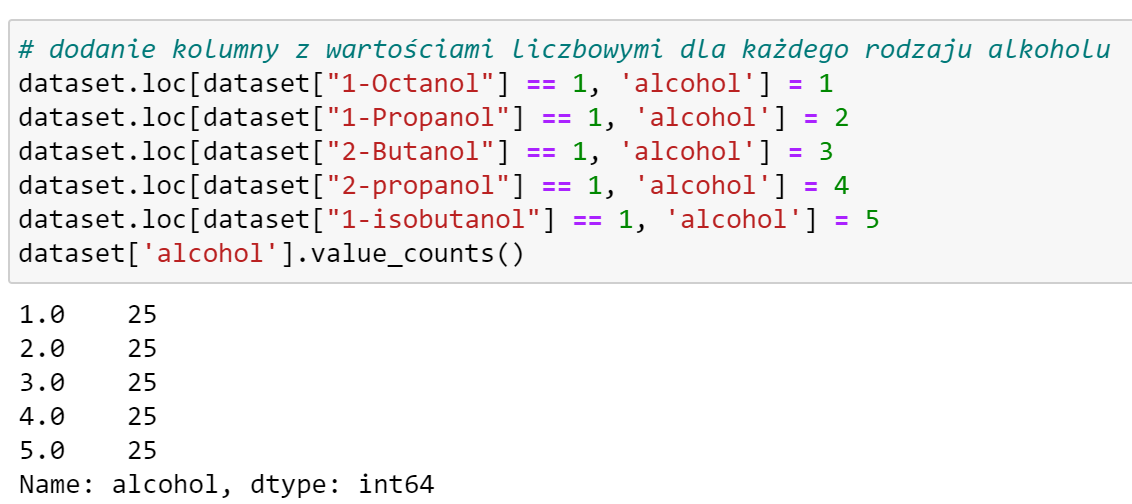
Obraz 11: Użyte biblioteki



Obraz 12: Załadowanie danych z plików

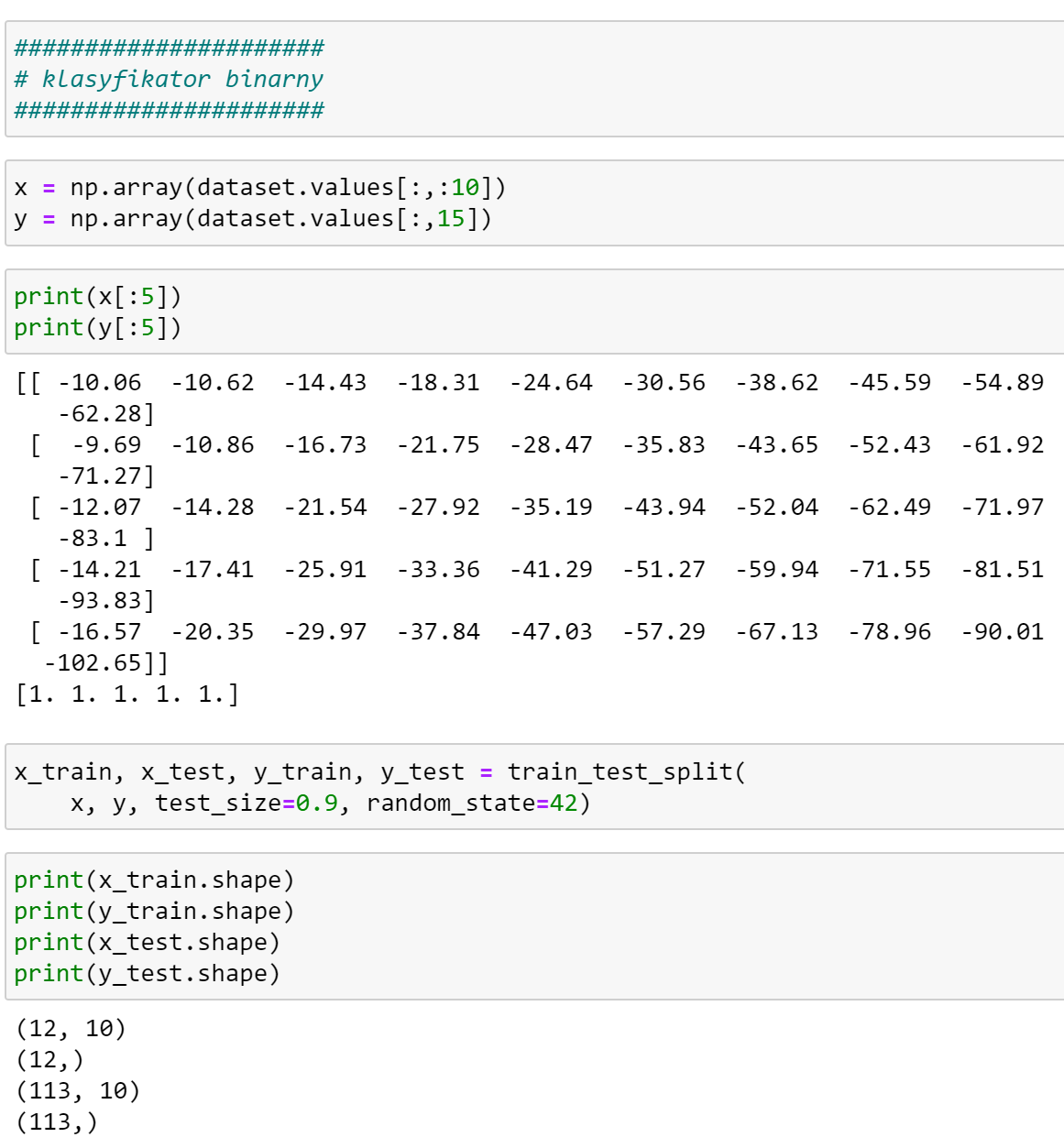


Obraz 13: Połączenie danych w jeden zestaw

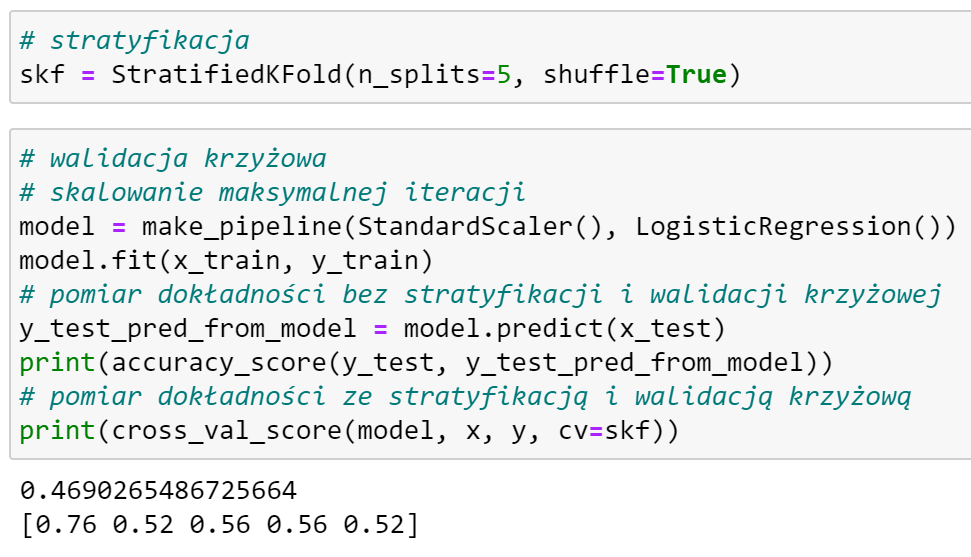


Obraz 14: Dodanie kolumny sumarycznej klasyfikującej rodzaj użytego alkoholu

### Klasyfikator: binarny



Obraz 15: Podział na zbiór trenujący i testujący

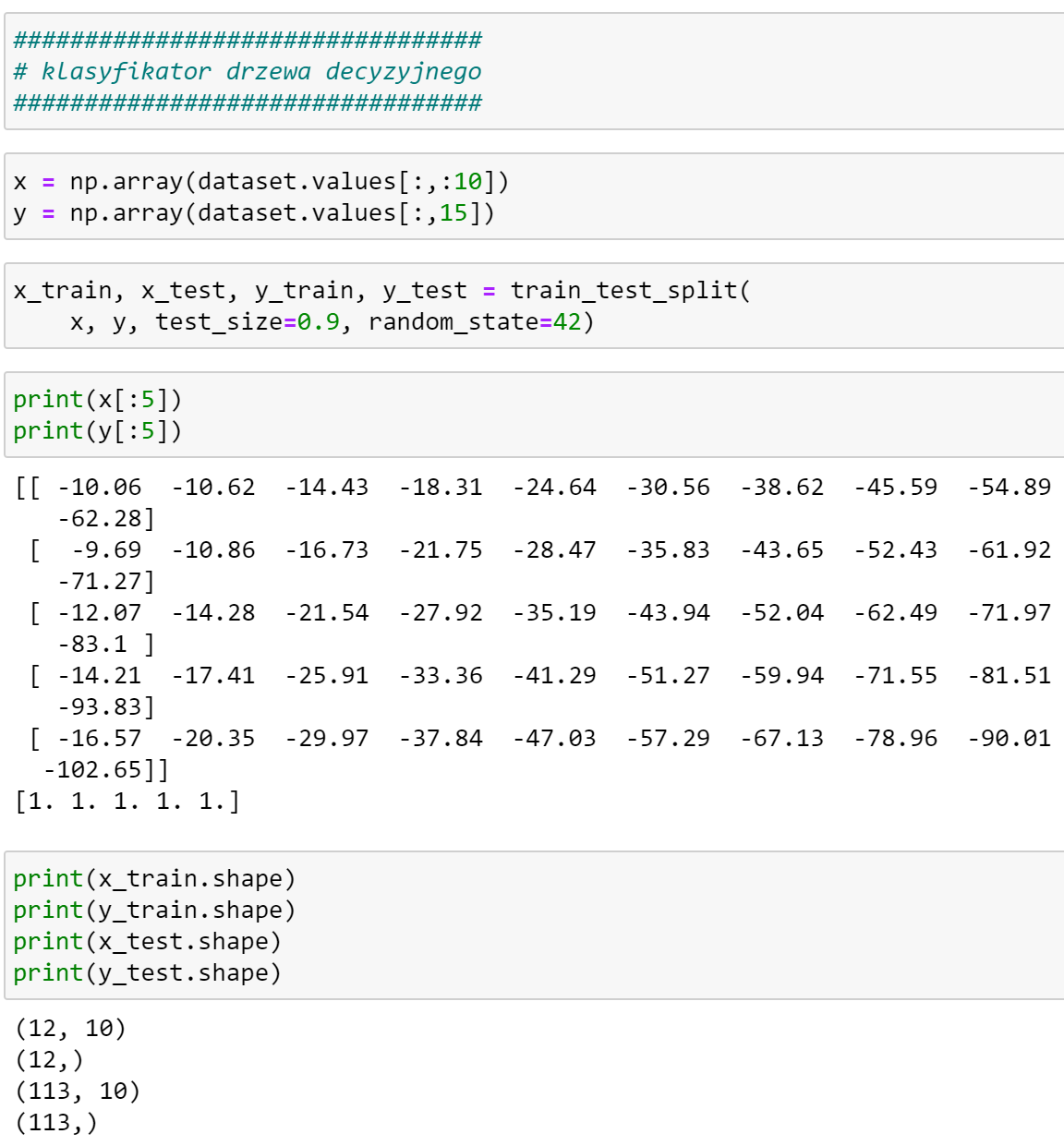


Obraz 16: Obliczenie dokładności modelu dla klasyfikatora binarnego

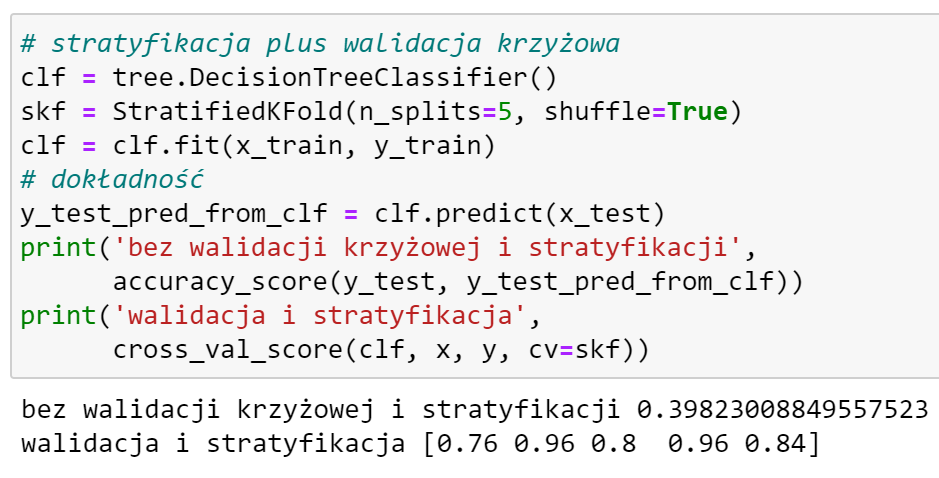
Pomimo faktu, że zbiór jest zbalansowany stratyfikacja w połączeniu z walidacją krzyżową wydają się podnosić dokładność modelu.

Nie zastosowano optymalizacji modelu ze względu na złożoność klasyfikatora.

### Klasyfikator: drzewo decyzyjne



Obraz 17: Podział na zbiór trenujący i testujący



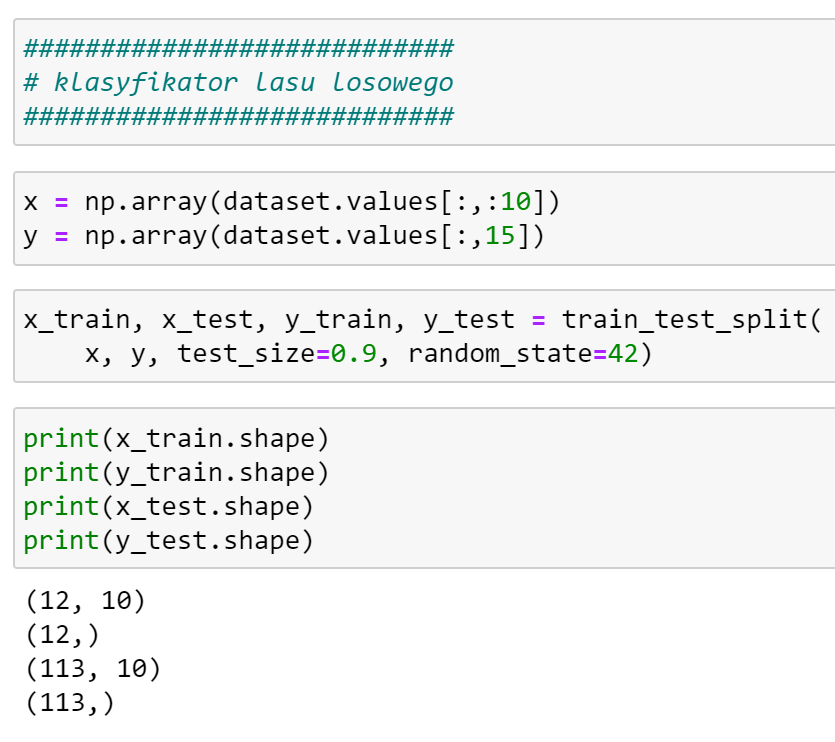
Obraz 18: Obliczenie dokładności modelu z klasyfikatorem drzewa decyzyjnego bez optymalizacji



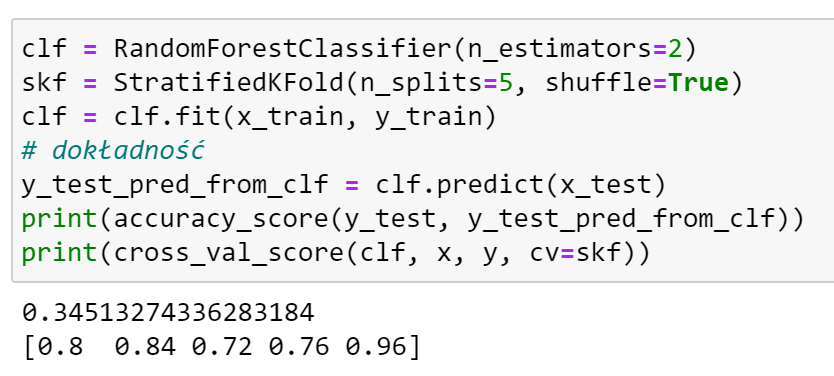
Obraz 19: Optymalizacja modelu i obliczenie dokładności zoptymalizowanego modelu z klasyfikatorem drzewa decyzyjnego

Pomimo faktu, że zbiór jest zbalansowany stratyfikacja w połączeniu z walidacją krzyżową wydają się podnosić dokładność modelu.

### Klasyfikator: las losowy



Obraz 20: Podział na zbiór trenujący i testujący



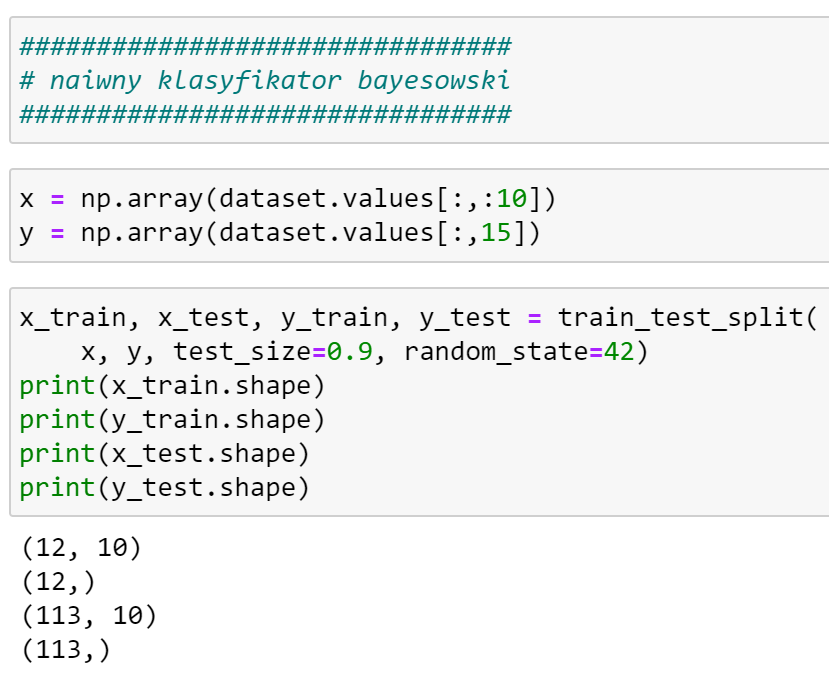
Obraz 21: Obliczenie dokładności modelu z klasyfikatorem lasu losowego bez optymalizacji



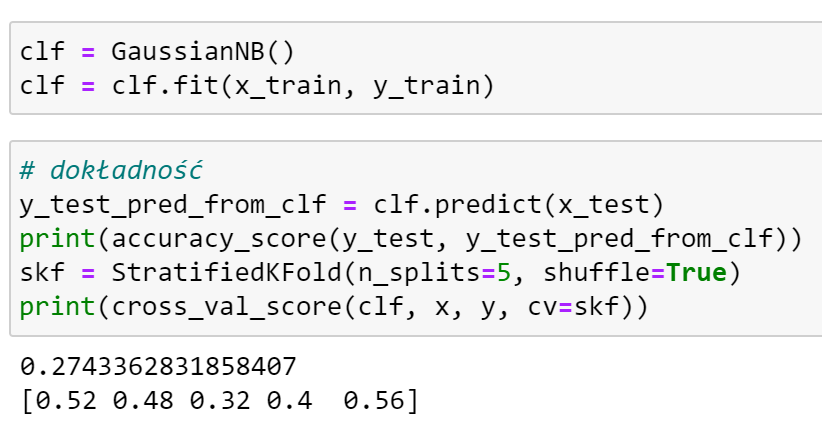
Obraz 22: Optymalizacja modelu i obliczenie dokładności zoptymalizowanego modelu z klasyfikatorem lasu losowego

Pomimo faktu, że zbiór jest zbalansowany stratyfikacja w połączeniu z walidacją krzyżową podnoszą dokładność modelu.

### Klasyfikator: naiwny bayesowski



Obraz 23: Podział na zbiór trenujący i testujący

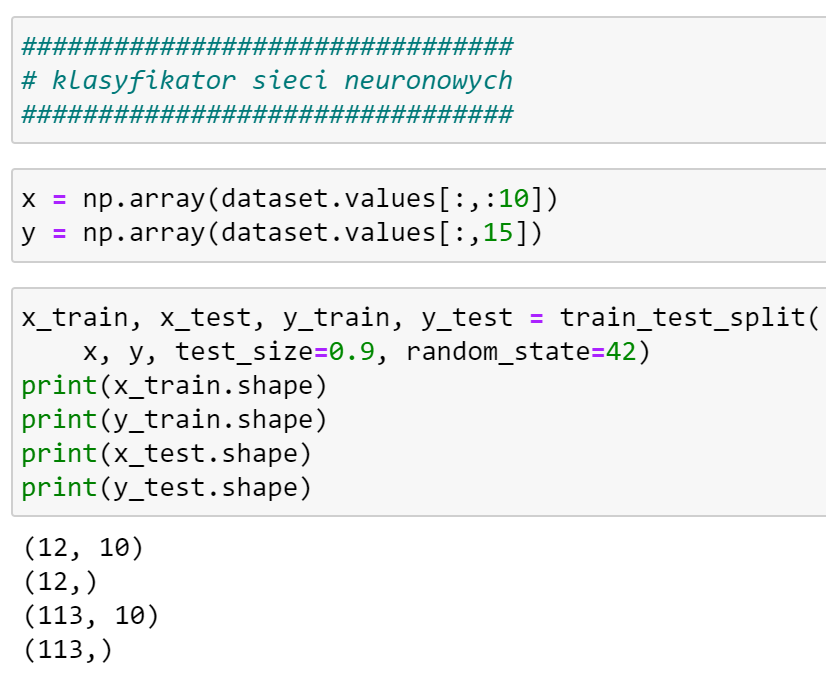


Obraz 24: Obliczenie dokładności modelu z naiwnym klasyfikatorem bayesowskim bez optymalizacji

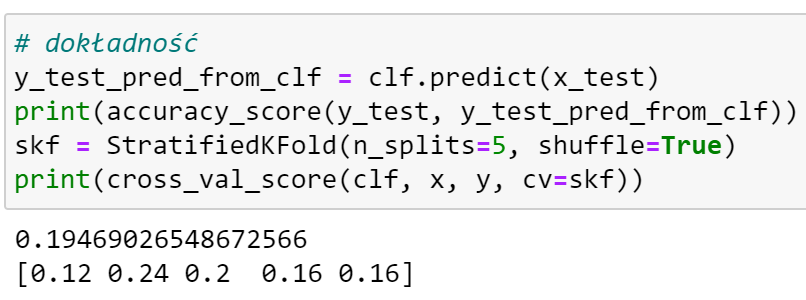
Pomimo faktu, że zbiór jest zbalansowany stratyfikacja w połączeniu z walidacją krzyżową wydają się podnosić dokładność modelu.

Nie zastosowano optymalizacji modelu ze względu na złożoność klasyfikatora.

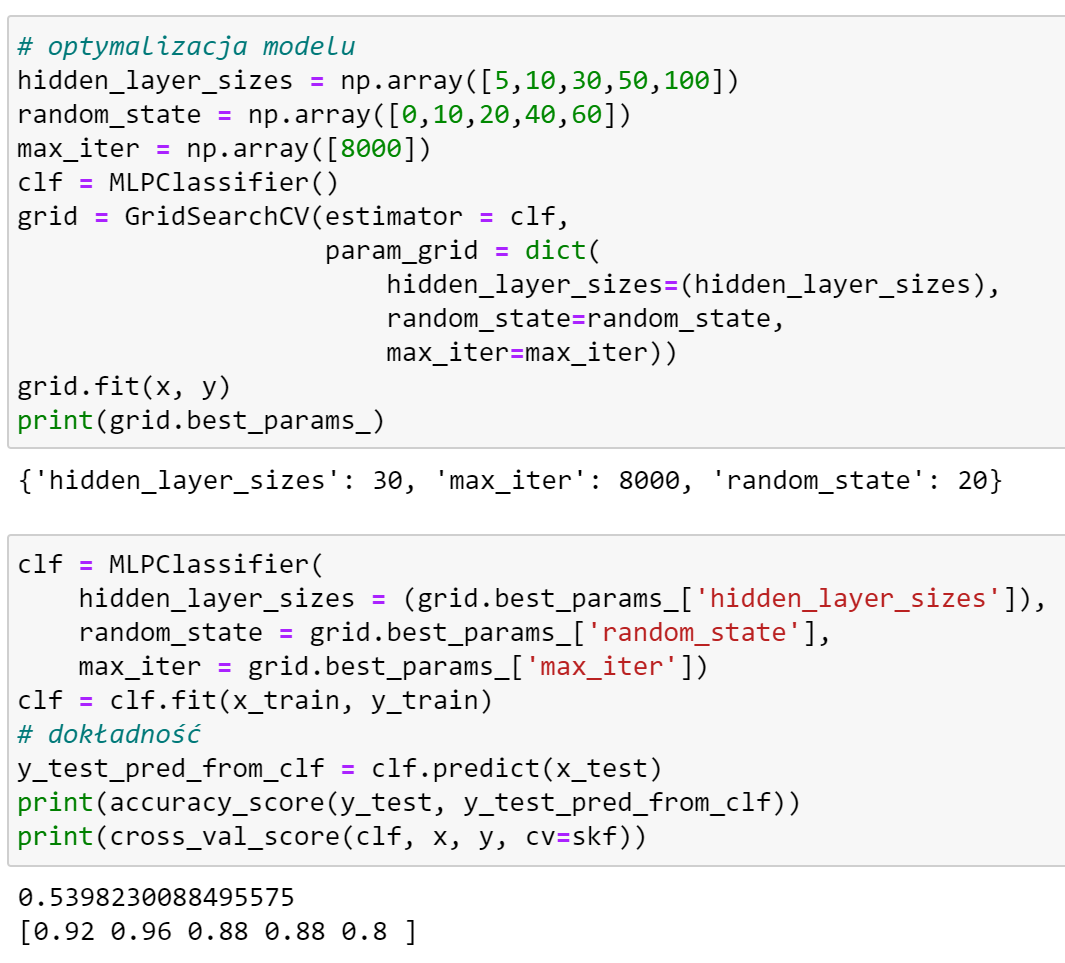
### Klasyfikator: sieci neuronowe



Obraz 25: Podział na zbiór trenujący i testujący



Obraz 26: Obliczenie dokładności modelu z klasyfikatorem sieci neuronowych bez optymalizacji



Obraz 27: Optymalizacja modelu i obliczenie dokładności zoptymalizowanego modelu z klasyfikatorem sieci neuronowych

Pomimo faktu, że zbiór jest zbalansowany stratyfikacja w połączeniu z walidacją krzyżową podnosi dokładność modelu.

## Podsumowanie i wnioski

W porównaniu dokładności modeli wzięto wyniki dla najlepiej dopasowanego modelu   
- z zastosowaniem grid\_search jeżeli było to możliwe oraz ze stratyfikacją i walidacją krzyżową, ponieważ pomiary wykazały, że dokładność każdego z modeli była wtedy większa.

Tabela 4: Zestawienie średnich dokładności najlepiej dopasowanych modeli dla różnych klasyfikatorów

|  |  |
| --- | --- |
| **Klasyfikator** | **Średnia dokładność modelu** |
| binarny | 0,560 |
| drzewa decyzyjnego | 0,896 |
| lasu losowego | 0,912 |
| naiwny bayesowski | 0,424 |
| sieci neuronowe | 0,920 |

Dla wybranego zestawu danych najlepiej poradziły sobie modele z klasyfikatorami: lasu losowego i sieci neuronowych, model z klasyfikatorem drzewa decyzyjnego również ma bardzo wysoką dokładność. Modele z klasyfikatorami: binarnym i naiwnym bayesowskim mają dokładność rzędu 50%, jednak warto zauważyć że podczas wykonywania ćwiczenia były to modele bez zastosowania optymalizacji grid\_search.

Oprócz wyniku dokładności warto również pamiętać o złożoności klasyfikatorów i czasu wykonywania obliczeń przez program. Klasyfikator sieci neuronowych jest najbardziej złożonym z wszystkich użytych w ćwiczeniu, najtrudniej było wybrać zakres danych podczas optymalizacji modelu i czas wykonywania obliczeń był zdecydowanie najdłuższy. Klasyfikatory: drzewa decyzyjnego i lasu losowego mają równie wysoką dokładność, a ich złożoność jest dużo mniejsza i wykonują się szybciej.

1. Opracowanie na podstawie analizy tongahancepel: <https://www.kaggle.com/tolgahancepel/qcm-sensor-alcohol-classification-using-keras/notebook>   
   oraz artykułu „*Classification of alcohols obtained by QCM sensors with different characteristics using ABC based neural network*” M. Fatih Adak, Peter Lieberzeit, Purim Jarujamrus, Nejat Yumusak (<https://www.sciencedirect.com/science/article/pii/S2215098619303337>) [↑](#footnote-ref-0)
2. the Qaurtz Crystal Microbalance - mikrowaga kwarcowa [↑](#footnote-ref-1)
3. Informacje pochodzą z Wikipedii: <https://pl.wikipedia.org/wiki/Mikrowaga_kwarcowa> [↑](#footnote-ref-2)
4. Więcej na Wikipedii: <https://pl.wikipedia.org/wiki/Elektroniczny_nos> [↑](#footnote-ref-3)
5. moleculary imprinted polymers - [↑](#footnote-ref-4)
6. Nanoparticles - [↑](#footnote-ref-5)