Probabilistyczne Uczenie Maszynowe Projekt 1

Jakub Dziwiński, Katarzyna Jabłońska, Dominika Kunc

Cel projektu

Celem projektu było dokonanie klasyfikacji stosowania danych używek przez osoby na podstawie cech demograficznych, osobowości oraz stosowania innych substancji.

Wybór docelowej używki użytej do klasyfikacji został dokonany po eksploracyjnej analizie danych.

Elementy projektu:

- Eksploracyjna analiza danych
- Modele mikstur rozkładów normalnych
- Naiwny Bayes

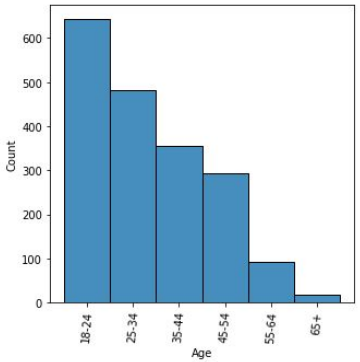
Zbiór danych

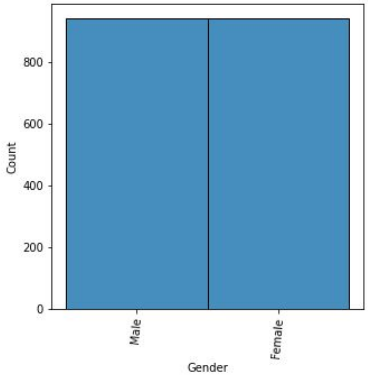
Zbiór danych rozważany w tym projekcie to zbiór ***Drug consumption*** dostępny na UCI Machine Learning Repository.

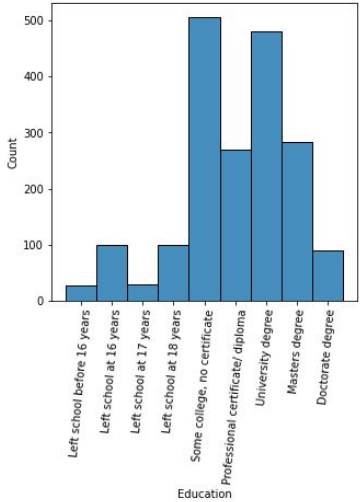
- 1885 rekordów
- 30 danych o użytkownikach:
  - dane demograficzne (wiek, płeć, poziom edukacji, kraj zamieszkania, pochodzenie etniczne)
  - dane o osobowości (neurotyzm, ekstrawersja, otwartość na doświadczenia, ugodowość, sumienność, impulsywność, poszukiwanie nowych doznań
  - dane dotyczące stosowania używek - 19 substancji (m.in. alkohol, kofeina, czekolada, marihuana, nikotyna, amfetamina, LSD, kokaina i inne).
    Stosowanie używek można było określić za pomocą 6 klas:
    - CL0 - Nigdy nieużywane,
    - CL1 - Używane ponad dekadę temu,
    - CL2 - Używane w ciągu ostatniej dekady,
    - CL3 - Używane w ciągu ostatniego roku,
    - CL4 - Używane w ciągu ostatniego miesiąca,
    - CL5 - Używane w ciągu ostatniego tygodnia,
    - CL6 - Używane w ciągu ostatniego dnia.

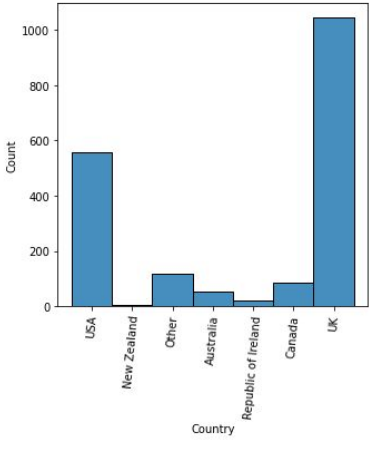
Eksploracyjna analiza danych

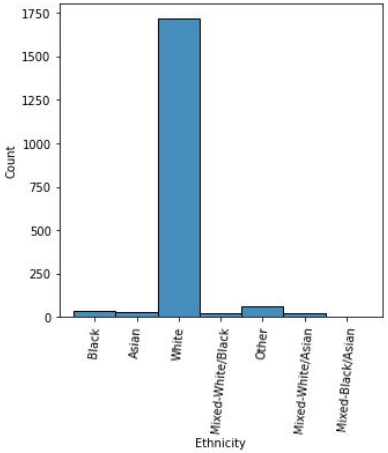
Dane demograficzne





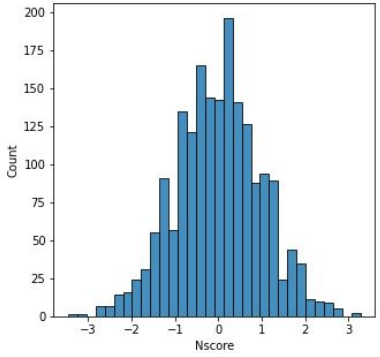


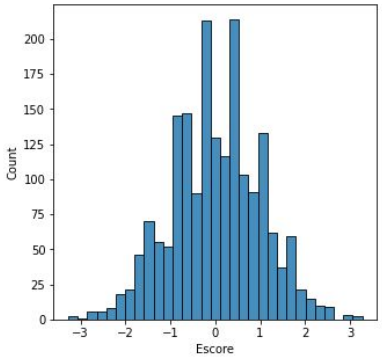


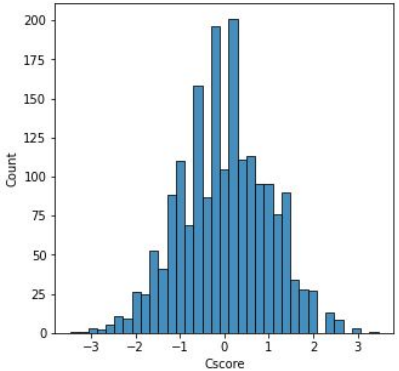


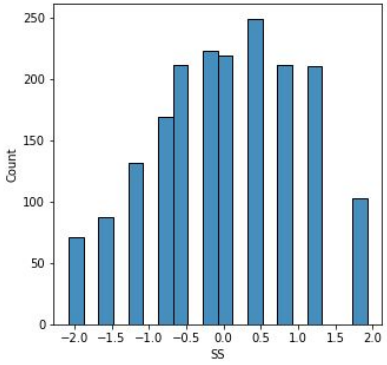
Eksploracyjna analiza danych

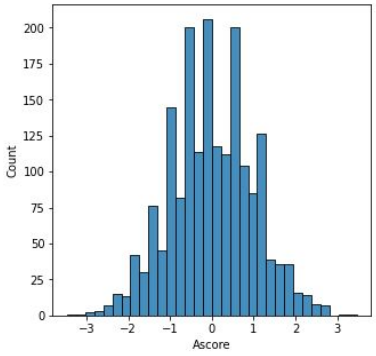
Dane dotyczące osobowości

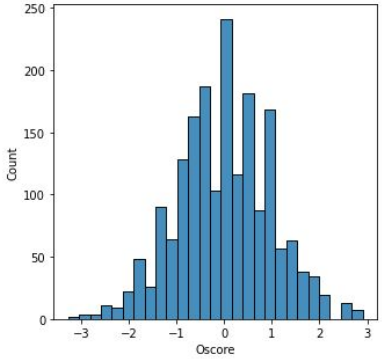


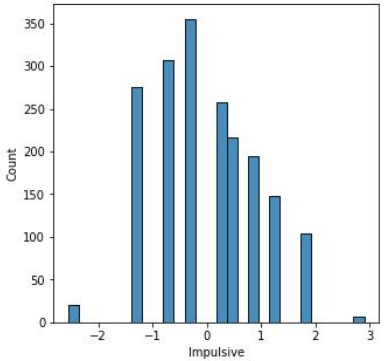






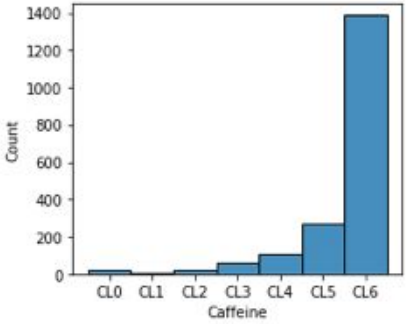


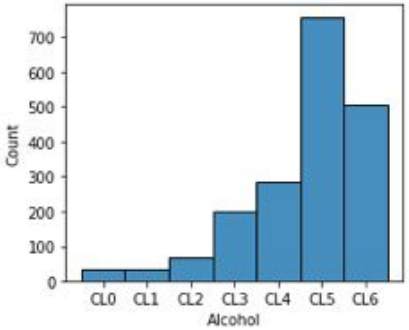


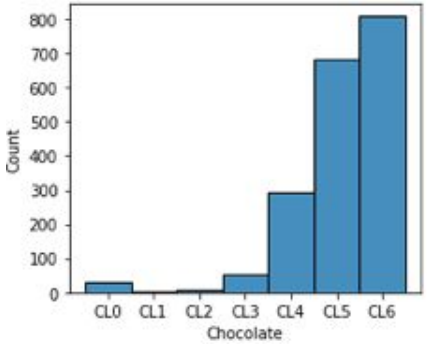


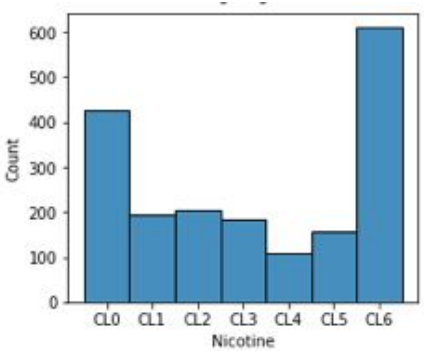
Eksploracyjna analiza danych

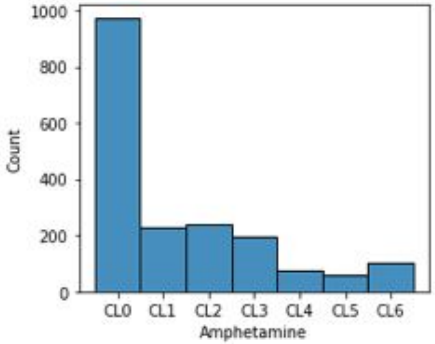
Dane dotyczące stosowania używek

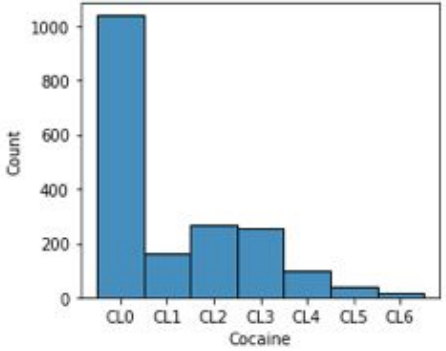


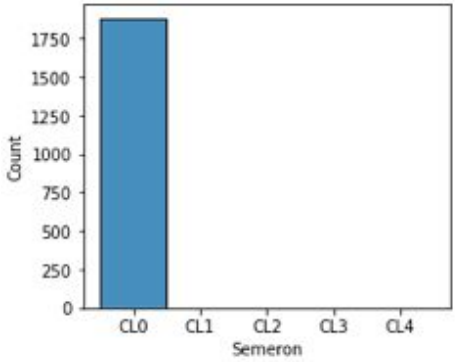












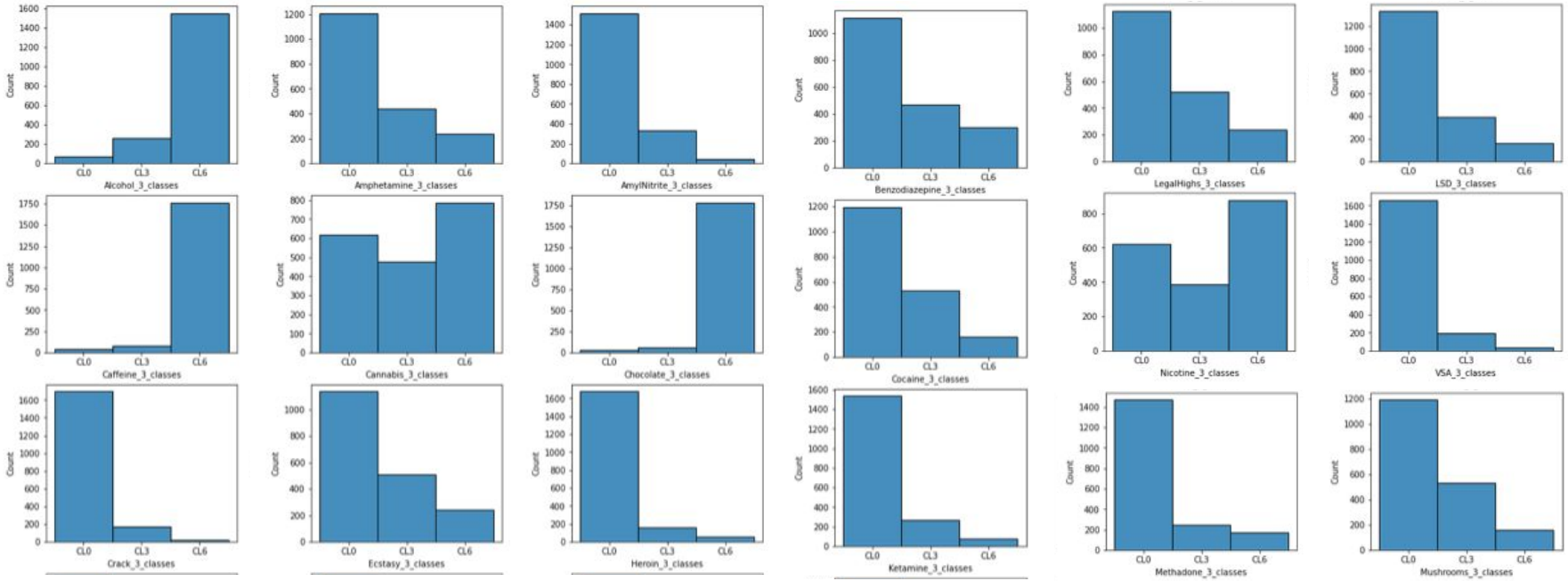
Eksploracyjna analiza danych

Zmniejszenie liczby klas dla stosowania używek:

- **CL0** - Nigdy nie używano - połączone CL0 (nigdy nie używano) oraz CL1 (używano ponad dekadę temu),
- **CL3** - Używane w ostatniej dekadzie - połączone CL2 (używane w ostatniej dekadzie) i CL3 (używane w ciągu ostatniego roku),
- **CL6** - używane w ostatnim miesiącu - połączone CL4 (używane w ostatnim miesiącu), CL5 (używane w ostatnim tygodniu) i CL6 (używane w ostatnim dniu).

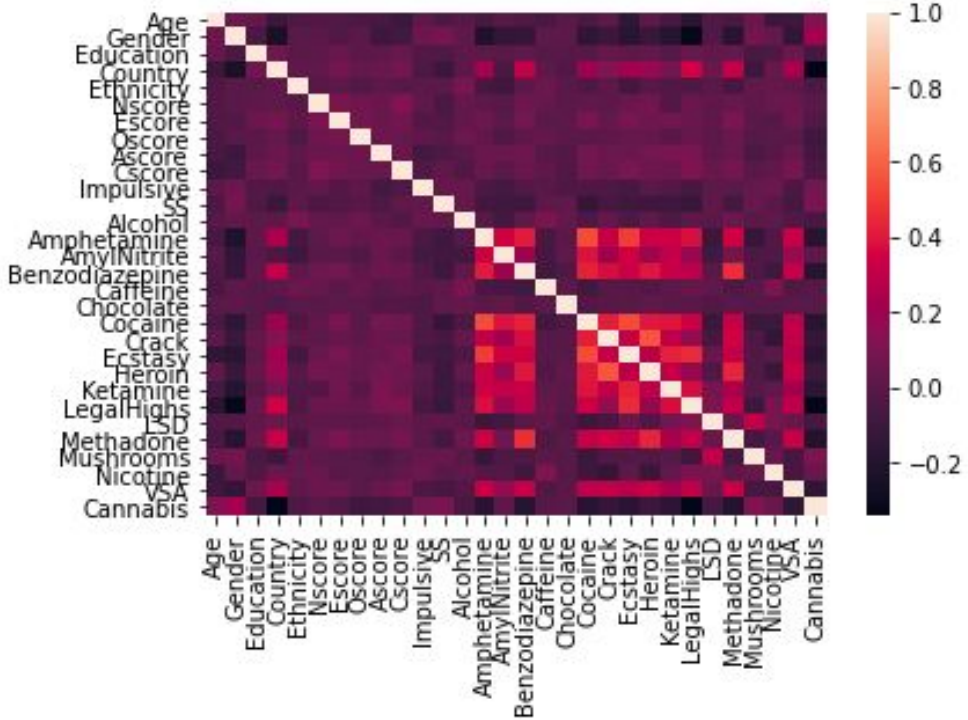
Eksploracyjna analiza danych

Dane dotyczące stosowania używek - zmniejszenie liczby klas



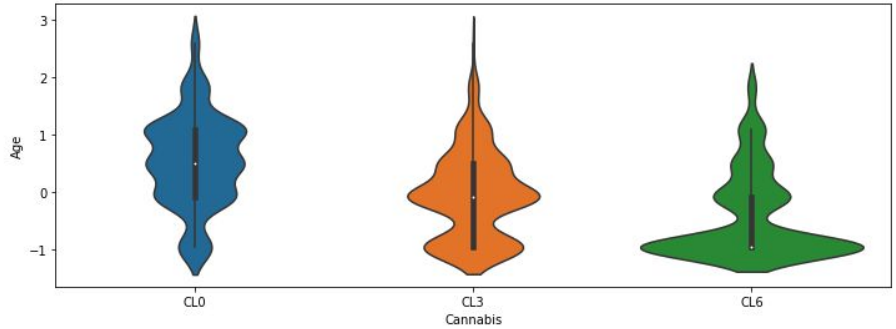
Eksploracyjna analiza danych

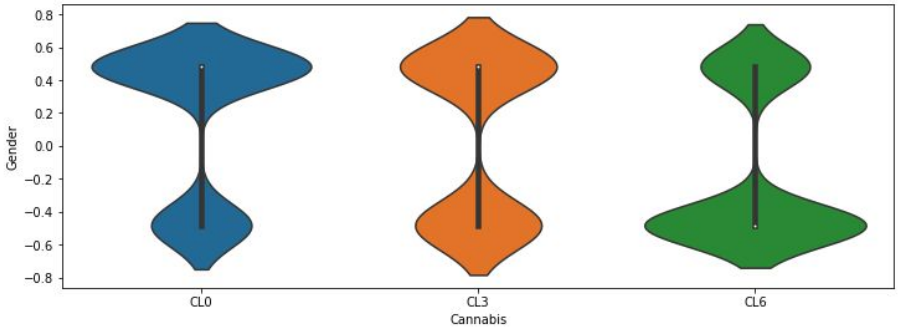
Korelacja Spearmana pomiędzy parami zmiennych

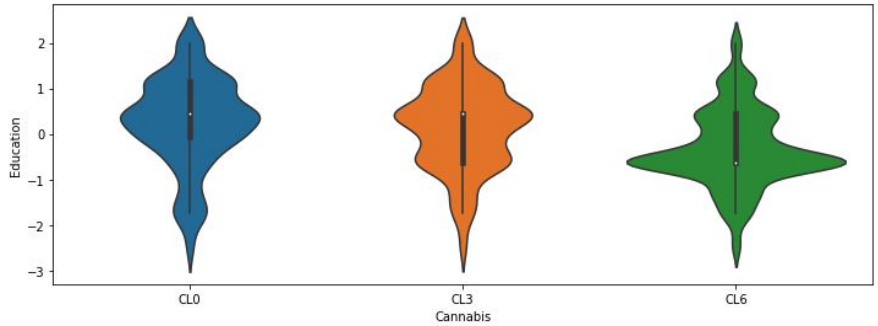


Eksploracyjna analiza danych

Zależności pomiędzy parami zmiennych - cechy demograficzne a stosowanie marihuany

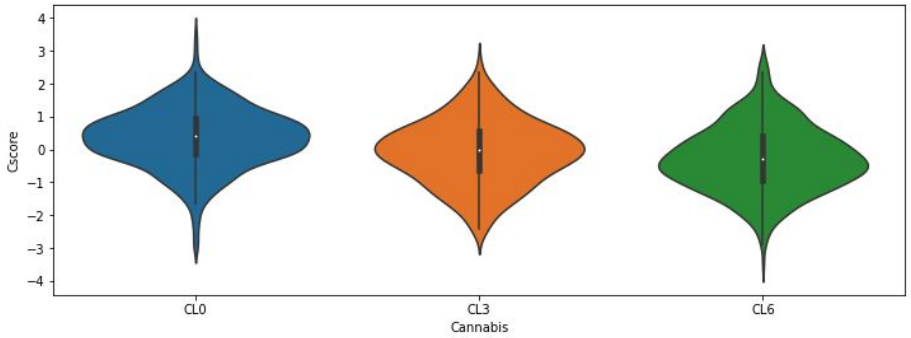


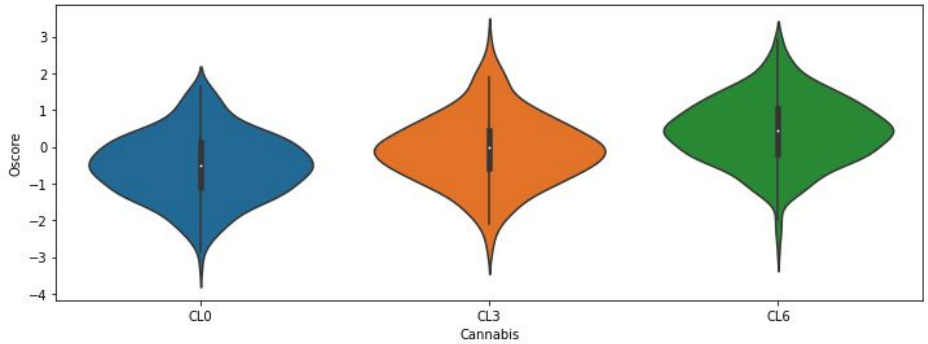


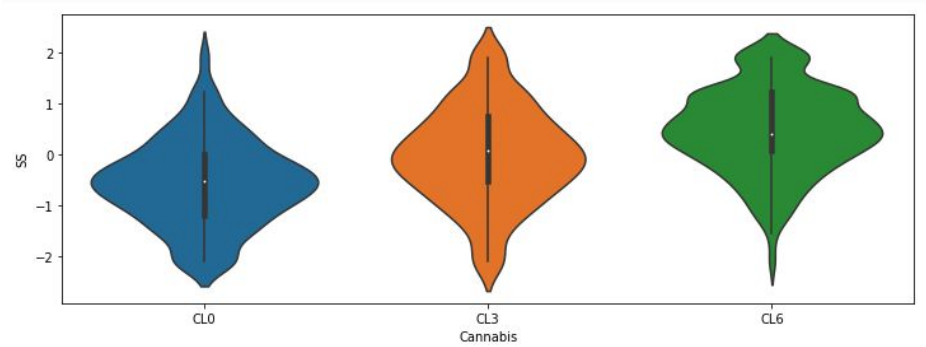


Eksploracyjna analiza danych

Zależności pomiędzy parami zmiennych - cechy osobowości a stosowanie marihuany

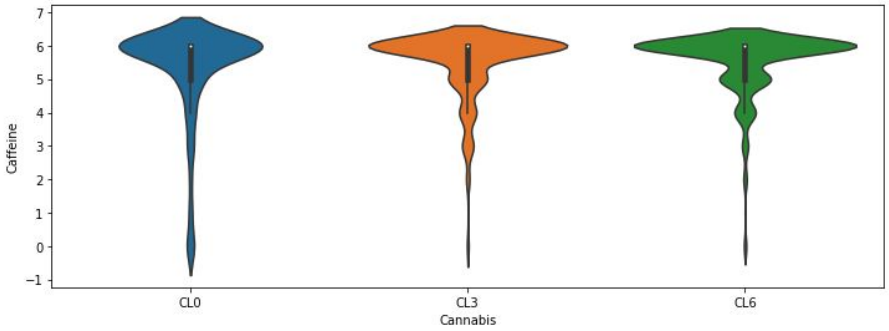


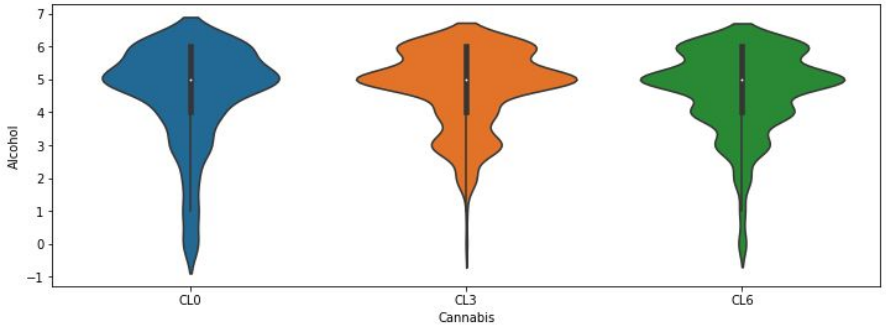


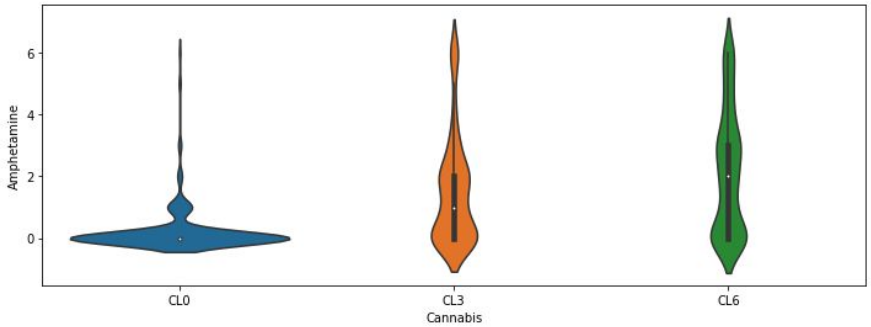


Eksploracyjna analiza danych

Zależności pomiędzy parami zmiennych - stosowanie używek a stosowanie marihuany







Przetwarzanie danych

- Zredukowanie liczby klas z 7 do 3
- Stratyfikowany podział zbioru na treningowy (70%) i testowy (30%)
- Usunięcie innych dotyczących używki “Semeron” ze względu na wyjątkowo silne niezbalansowanie
- Standaryzacja zbiorów (StandardScaler z biblioteki sci-kit learn)

Naiwny Bayes

Badane modele:

1. z biblioteki Sklearn:
   a. GaussianNB
   b. BernoulliNB
   c. MultinomialNB
   d. ComplementNB
   e. CategoricalNB

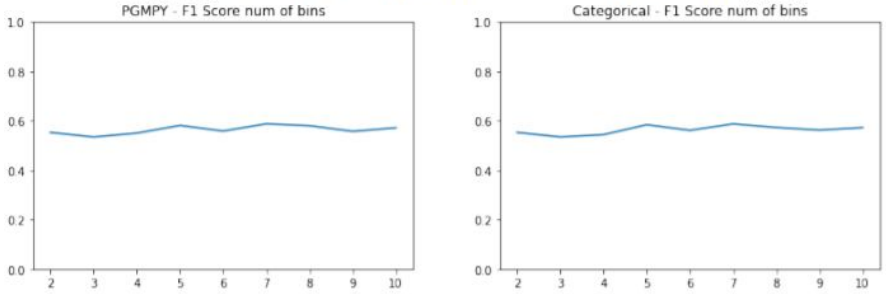
2. implementacja z wykorzystaniem biblioteki **pyro**
3. implementacja z wykorzystaniem biblioteki **pgmpy**

Zakres badań

1. Przyjmowane dane (wszystkie cechy lub tylko te związane z człowiekiem)
2. Parametr *K* dla dyskretyzatora *KBins* w przypadku klasyfikatorów potrzebujących dyskretnych danych
3. Liczba epok w przypadku klasyfikatora z biblioteki **pyro**

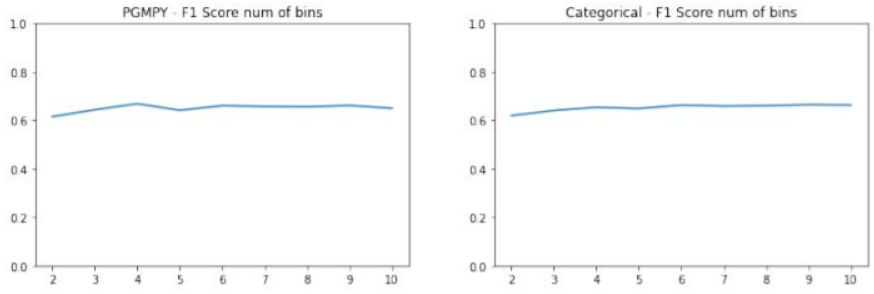
Zbiór opisujący spożycie marihuany – KBins

Cechy związane z człowiekiem



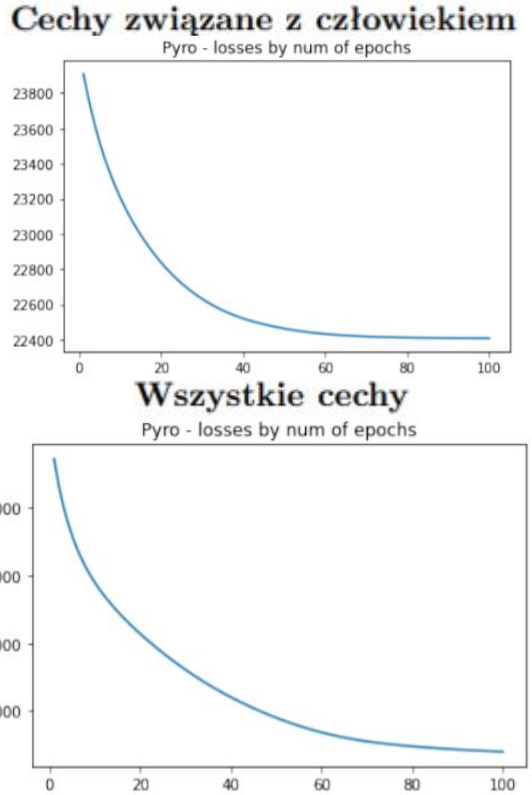
(a) Model *pgmpy* (b) CategoricalNB

Wszystkie cechy

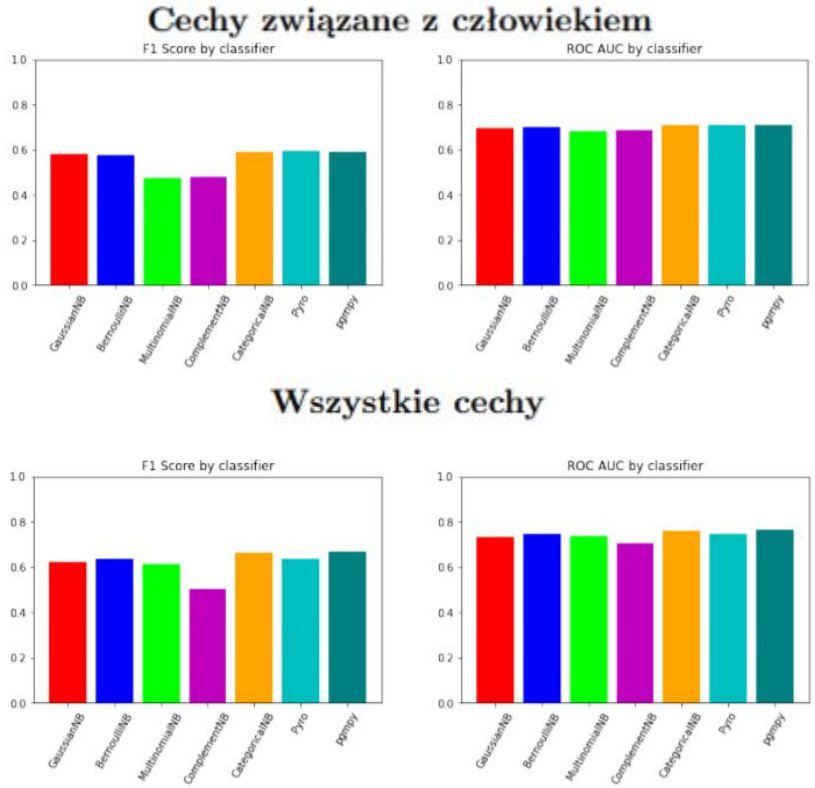


(a) Model *pgmpy* (b) CategoricalNB

Zbiór opisujący spożycie marihuany – loss



Zbiór opisujący spożycie marihuany – metryki



Zbiór opisujący spożycie marihuany – najlepsze wyniki

	precision	recall	f1-score	support
0	0.64	0.74	0.69	186
1	0.37	0.31	0.34	143
2	0.76	0.75	0.76	237
accuracy			0.64	566
macro avg	0.59	0.60	0.59	566
weighted avg	0.63	0.64	0.63	566

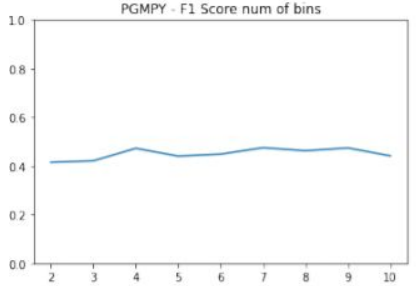
(a) Cechy związane z człowiekiem - model *pyro*

	precision	recall	f1-score	support
0	0.71	0.91	0.80	186
1	0.50	0.41	0.45	143
2	0.81	0.71	0.76	237
accuracy			0.70	566
macro avg	0.67	0.68	0.67	566
weighted avg	0.70	0.70	0.69	566

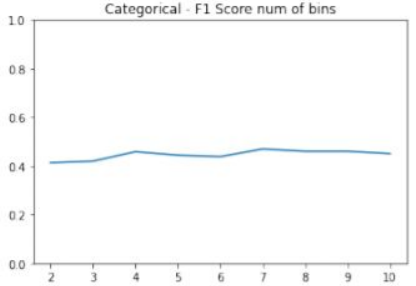
(b) Wszystkie cechy - model *pgmpy*

Zbiór opisujący spożycie nikotyny – KBins

Cechy związane z człowiekiem

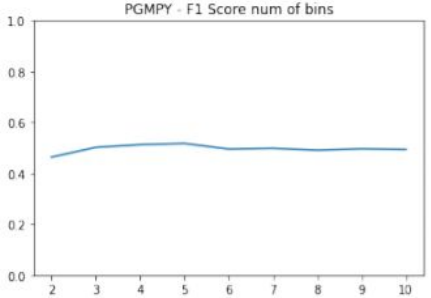


(a) Model *pgmpy*

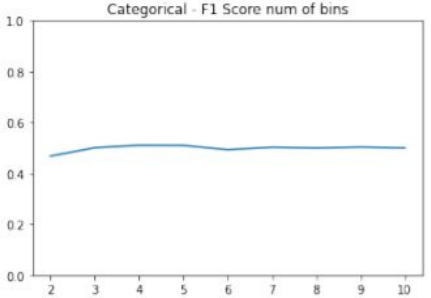


(b) CategoricalNB

Wszystkie cechy



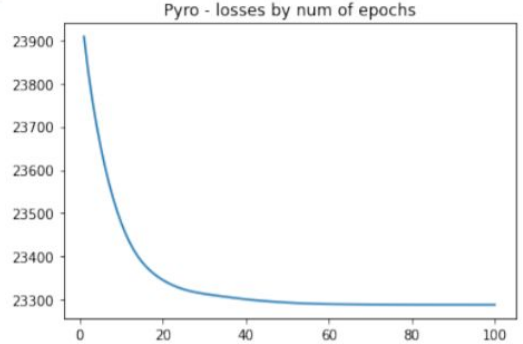
(a) Model *pgmpy*



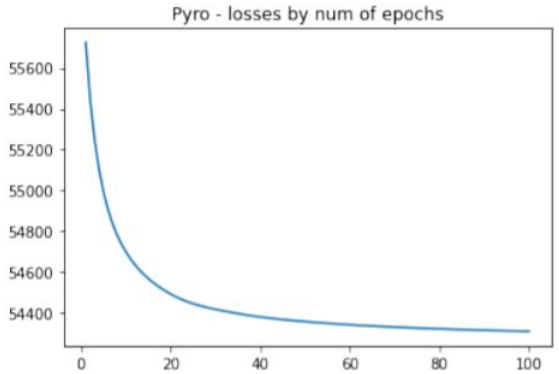
(b) CategoricalNB

Zbiór opisujący spożycie marihuany – loss

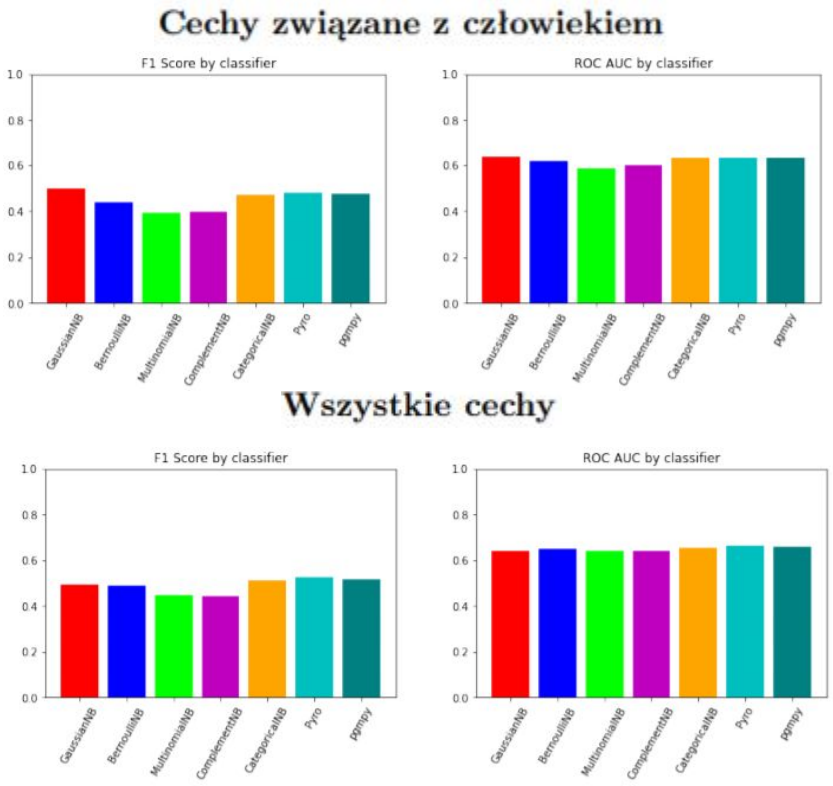
Cechy związane z człowiekiem



Wszystkie cechy



Zbiór opisujący spożycie marihuany – metryki



Zbiór opisujący spożycie nikotyny- najlepsze wyniki

	precision	recall	f1-score	support
0	0.58	0.66	0.61	186
1	0.28	0.23	0.25	117
2	0.63	0.62	0.62	263
accuracy			0.55	566
macro avg	0.50	0.50	0.50	566
weighted avg	0.54	0.55	0.54	566

(a) Cechy związane z człowiekiem - model *GaussianNB*

	precision	recall	f1-score	support
0	0.62	0.73	0.67	186
1	0.28	0.19	0.22	117
2	0.68	0.69	0.68	263
accuracy			0.60	566
macro avg	0.52	0.54	0.53	566
weighted avg	0.58	0.60	0.58	566

(b) Wszystkie cechy - model *pyro*

Modele mikstur rozkładów normalnych

Badane modele:

1. z biblioteki Sklearn,
2. implementacja własna

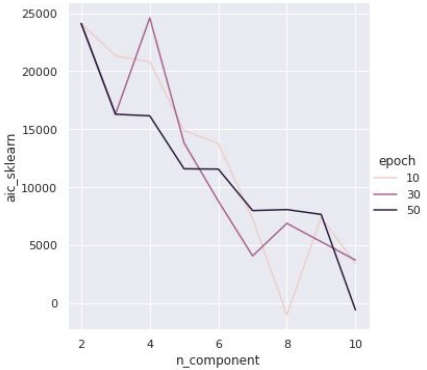
Plan badań

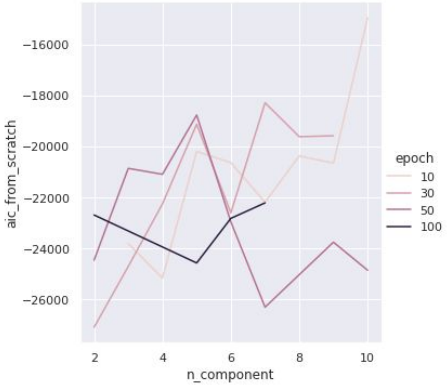
1. Wybór najlepszych hiperparametrów przy użyciu miar AIC oraz BIC
2. Wizualizacja najlepszych wyników - wnioski z klasyfikacji
3. Stworzenie metryk klasyfikacji na podstawie wizualizacji - porównanie wyników

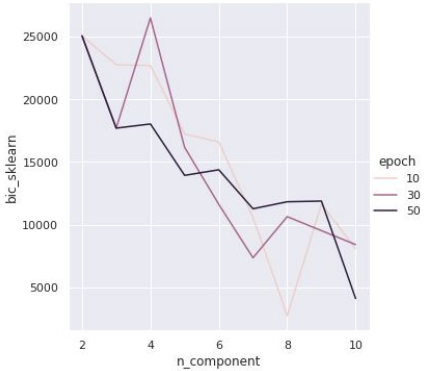
Testowane hiperparametry:

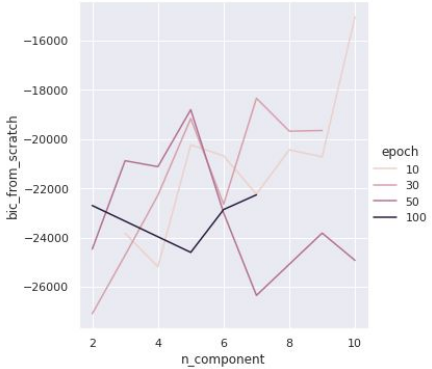
1. liczba komponentów [2, 3, 4, 5, 6, 7, 8, 9, 10]
2. liczba maksymalnych iteracji [10, 30, 50, 100]

Zbiór opisujący spożycie marihuany



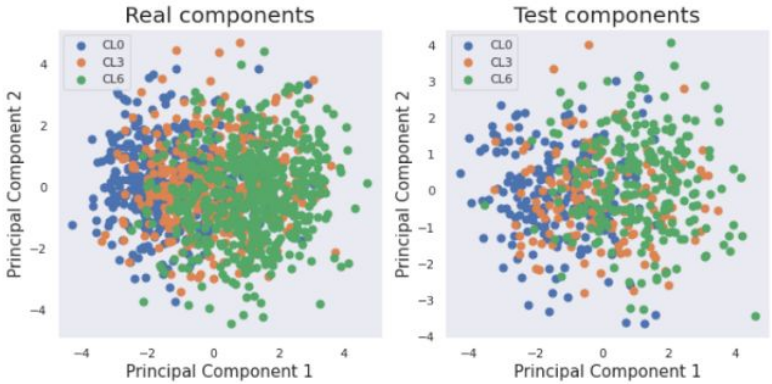




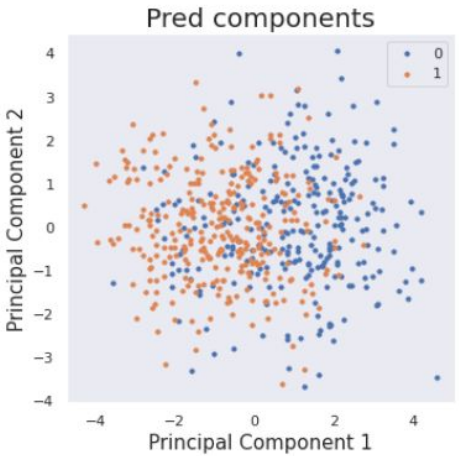


Sklearn

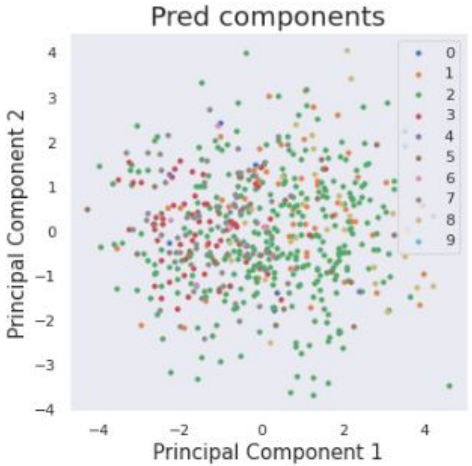
Implementacja własna



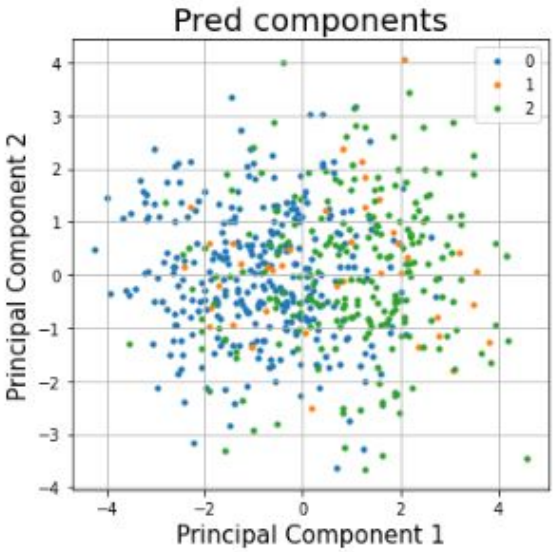
Prawdziwy rozkład klastrów dla zbioru treningowego oraz testowego



Sklearn

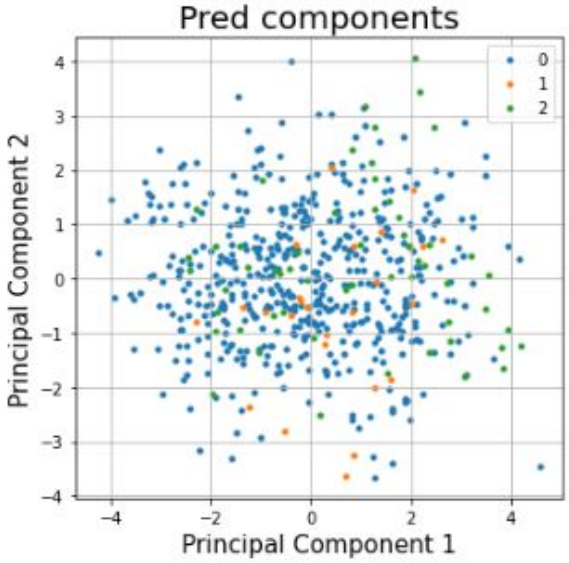


Implementacja własna



	precision	recall	f1-score	support
0	0.48	0.78	0.60	186
1	0.20	0.06	0.09	143
2	0.66	0.62	0.64	237
accuracy			0.53	566
macro avg	0.45	0.49	0.44	566
weighted avg	0.48	0.53	0.49	566

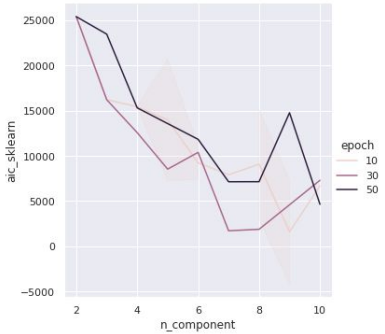
Sklearn

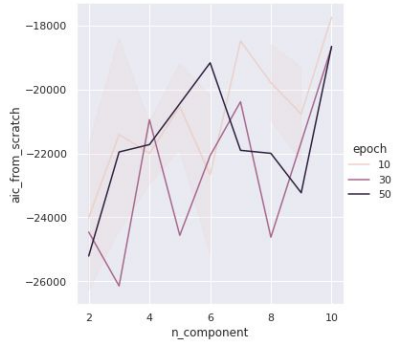


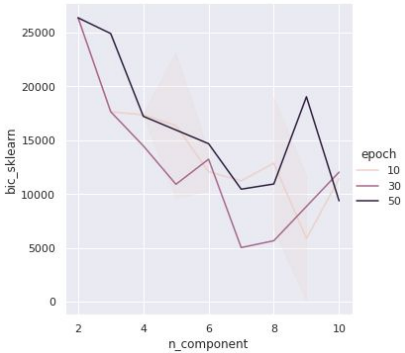
	precision	recall	f1-score	support
0	0.34	0.89	0.50	186
1	0.31	0.06	0.09	143
2	0.54	0.13	0.21	237
accuracy			0.36	566
macro avg	0.40	0.36	0.27	566
weighted avg	0.42	0.36	0.28	566

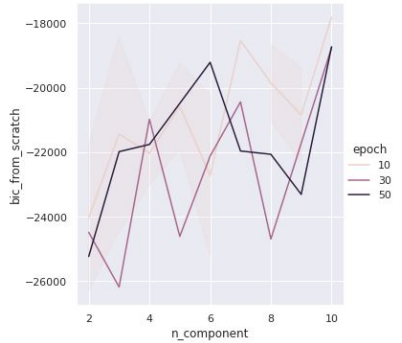
Implementacja własna

Zbiór opisujący spożycie nikotyny



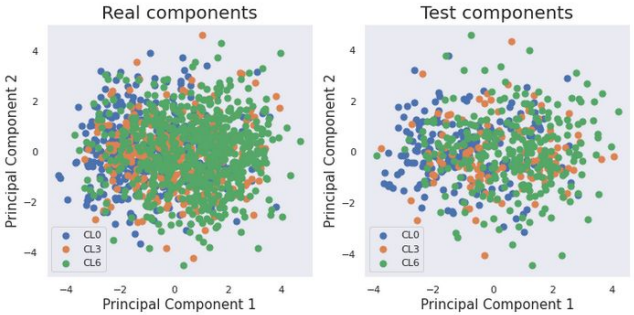


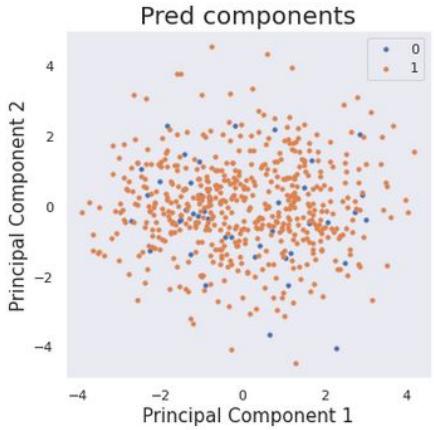


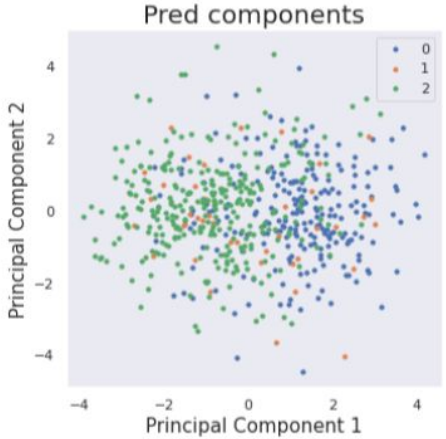


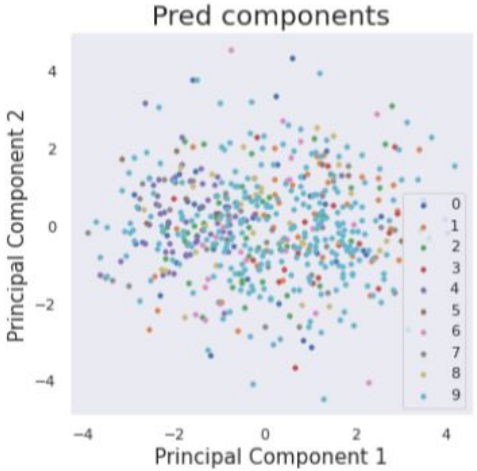
Sklearn

Implementacja własna



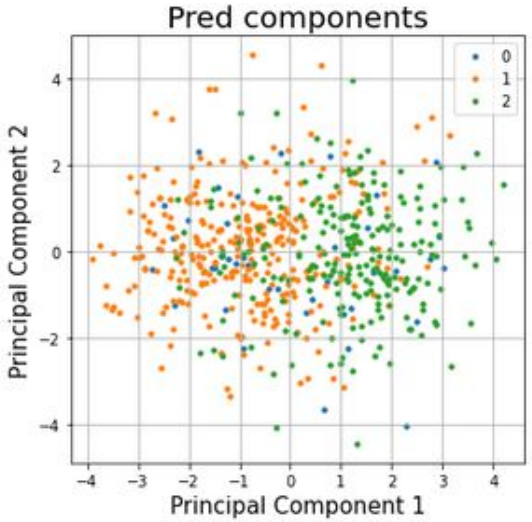






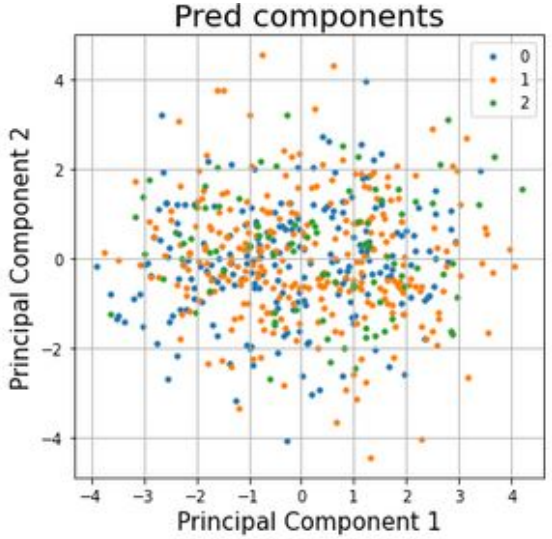
Sklearn

Implementacja własna



	precision	recall	f1-score	support
0	0.21	0.08	0.11	117
1	0.43	0.69	0.53	186
2	0.55	0.47	0.51	263
accuracy			0.46	566
macro avg	0.40	0.41	0.38	566
weighted avg	0.44	0.46	0.43	566

Sklearn



	precision	recall	f1-score	support
0	0.35	0.39	0.37	186
1	0.45	0.47	0.46	263
2	0.11	0.09	0.10	117
accuracy			0.36	566
macro avg	0.31	0.31	0.31	566
weighted avg	0.35	0.36	0.36	566

Implementacja własna

Wnioski

- Analiza eksploracyjna potwierdziła pierwotne hipotezy o stosowaniu różnych używek
- Stosowanie pewnych używek może mieć wpływ na stosowanie innych (macierz korelacji)
- Lepsze rezultaty w klasyfikacji uzyskał Naiwny Bayes
- Model mikstur rozkładów normalnych nie zadziałały dla tego zadania, jednak mimo wszystko dość dobrze odwzorowywały poszczególne zależności

Dziękujemy za uwagę!