Probabilistyczne Uczenie Maszynowe
Projekt 1

Jakub Dziwiński, Katarzyna Jabłońska, Dominika Kunc

Kwiecień 2021

Spis treści

1 Wstęp

Niniejszy raport dotyczy realizacji pierwszego projektu z Probabilistycznego Uczenia Maszynowego. Projekt zakłada dokonanie eksploracyjnej analizy danych i ich odpowiedniego przetworzenia, zaimplementowania prostego modelu probabilistycznego, jak i probabilistycznego modelu grafowego. Każdy z elementów składowych projektu, wraz z eksperymentami oraz uzyskanymi wynikami jest opisany w kolejnych rozdziałach.

2 Eksploracyjna analiza danych

2.1 Zbiór danych

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

Zbiór ten zawiera 1885 rekordów - odpowiedzi respondentów. Dla każdego z nich wyróżniamy dwa rodzaje danych:

- Dane demograficzne oraz atrybuty związane z osobowością,
- Dane dotyczące stosowania różnych używek.

W sekcji dotyczącej danych demograficznych zawarte jest 5 pozycji:

- Wiek (*Age*) - 6 przedziałów wiekowych (18-24, 25-34, 35-44, 45-54, 55-64, 65+)
- Płeć (*Gender*) - 2 wartości (Kobieta (*Female*) i Mężczyzna (*Male*))
- Poziom edukacji (*Education*) - 9 wartości:
  - Porzucił szkołę przed 16 rokiem życia (*Left school before 16 years*),
  - Porzucił szkołę w 16 roku życia *Left school at 16 years*),
  - Porzucił szkołę w 17 roku życia (*Left school at 17 years*),
  - Porzucił szkołę w 18 roku życia (*Left school at 18 years*),
  - Pewien uniwersytet, lecz nieukończony (*Some college or university - no certificate or degree*),
  - Profesjonalny certyfikat (*Professional certificate/ diploma*),
  - Dyplom uniwersytetu (*University degree*),
  - Dyplom magistra (*Masters degree*),
  - Doktorat (*Doctorate degree*).

- Kraj, w którym obecnie mieszka badana osoba (*Country*) - 7 wartości:
  - Australia (*Australia*),
  - Kanada (*Canada*),
  - Nowa Zelandia (*New Zealand*),
  - Inne (*Other*),
  - Republika Irlandii (*Republic of Ireland*),
  - Wielka Brytania (*UK*),
  - Stany Zjednoczone (*USA*).
- Pochodzenie etniczne (*Ethnicity*) - 7 wartości:
  - Azjatyckie (*Asian*),
  - Afroamerykańskie (*Black*),
  - Mieszane afroamerykańsko-azjatyckie (*Mixed-Black/Asian*),
  - Mieszane biało-azjatyckie (*Mixed-White/Asian*),
  - Mieszane biało-czarne(*Mixed-White/Black*),
  - Inne (*Other*),
  - Białe (*White*).

Dane dotyczące osobowości pozyskane zostały przy użyciu kwestionariuszy NEO-FFI-R, BIS-11 oraz ImpSS.

NEO-FFI-R jest to kwestionariusz służący do diagnozy cech osobowości - neurotyczności, ekstrawersji, otwartości na doświadczenie, ugodowość i sumienność. Polega on na wypełnieniu 60 twierdzeń samoopisowych. Uzyskane wyniki pozwalają na pełny opis osobowości badanego, a autorzy kwestionariusza polecają to narzędzie w szczególności do badań naukowych.

BIS-11 to skala szeroko stosowana miara impulsywności. Składa się ona z 30 twierdzeń opisujących popularne impulsywne lub nieimpulsywne zachowania i preferencje, które zostają ocenione w 4 punktowej skali (rzadko/nigdy, okazjonalnie, często, prawie zawsze/zawsze).

Z kolei skala ImpSS to 19 pytań prawda-fałsz, która przypisuje różnorodne cechy charakterystyczne i zachowania powiązane z impulsywnością i poszukiwaniem nowych doznań, a wynik uzyskuje się poprzez zsumowanie punktów za poszczególne odpowiedzi, stąd wartości znajdują się w przedziale 0-19.

Na podstawie powyższych kwestionariuszy uzyskanych zostało 7 cech:

- NScore (NEO-FFI-R) - neurotyzm (49 unikatowych wartości),
- Escore (NEO-FFI-R) - ekstrawersja (42 unikatowych wartości),
- Oscore (NEO-FFI-R) - otwartość na doświadczenia (35 unikatowych wartości),
- Ascore (NEO-FFI-R) - ugodowość (41 unikatowych wartości),
- Cscore (NEO-FFI-R) - sumienność (41 unikatowych wartości),
- Impulsive (BIS-11) - impulsywność (10 unikatowych wartości),
- SS (*Sensation Seeking* (ImpSS) - poszukiwanie nowych doznań (11 unikatowych wartości).

Wszystkie dotychczas wymienione cechy zostały skwantyfikowane - każda zmienna była traktowana jako kategoryczna (nawet wartości punktów w testach psychologicznych) i została zrzutowana na liczbę rzeczywistą.

W przypadku danych dotyczących stosowania używek rozważono 19 substancji:

- Alkohol (*Alcohol*),

- Amfetamina (*Amphetamine*),
- Azotyn Amylu (*Amyl Nitrite*),
- Benzodiazepina (*Benzodiazepine*),
- Kofeina (*Caffeine*),
- Marihuana (*Cannabis*),
- Czekolada (*Chocolate*),
- Kokaina (*Cocaine*),
- Krak (*Crack*),
- Ekstazy (*Ecstasy*),
- Heroina (*Heroin*),
- Ketamina (*Ketamine*),
- Legalne dopalacze (*Legal Highs*),
- LSD,
- Metadon (*Methadone*),
- Grzybki halucynogenne (*Mushrooms*),
- Nikotyna (*Nicotine*),
- Semeron,
- Wziewne środki odurzające (*VSA*).

Każdy z badanych ocenił każdą z powyższych używek pod względem częstotliwości używania. Oznacza to, że każda z nich mogła zostać sklasyfikowana do 1 z 7 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.

2.2 Zakresy wartości i rozkłady cech

Jak zostało już wcześniej wspomniane, każda z cech, która nie dotyczyła stosowania używek została skwantyfikowana. Z tego powodu oprócz liczby unikatowych wartości w tabeli 1 przedstawione zostały również wartości minimalne, maksymalne, średnie oraz odchylenia standardowe.

Cecha	Liczba unikalnych wartości	Min	Max	Średnia	Std
Wiek	6	-0.95197	2.59171	0.03461	0.87836
Płeć	2	-0.482460	0.482460	-0.000256	0.482588
Edukacja	9	-2.43591	1.98437	-0.0038	0.9501
Kraj zamieszkania	7	-0.57009	0.96082	0.3555	0.7003
Pochodzenie etniczne	7	-1.10702	1.90725	-0.3096	0.1662
Nscore	49	-3.46436	3.27393	0.00005	0.9981
Escore	42	-3.27393	3.27393	-0.0002	0.9974
Oscore	35	-3.27393	2.90161	-0.0005	0.9962
Ascore	41	-3.46436	3.46436	-0.0002	0.9974
Cscore	41	-3.46436	3.46436	-0.0004	0.99752
Impulsywność	10	-2.55524	2.90161	0.0072	0.9544
Poszukiwanie doznań	11	-2.07848	1.92173	-0.0033	0.9637

Tablica 1: Opis statystyczny cech związanych z danymi demograficznymi oraz osobowością badanych.

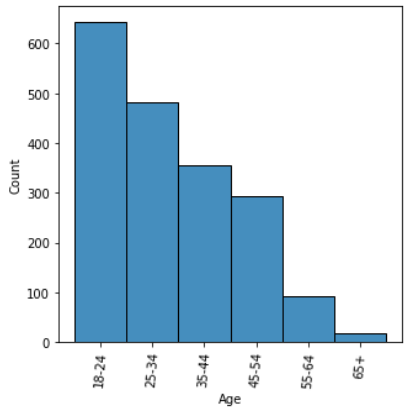
Na rysunku 1 przedstawione zostały rozkłady dla cech demograficznych. Z załączonych wykresów możemy odczytać informacje o badanych osobach, które mogą mieć spore znaczenie w interpretacji wyników dotyczących stosowania używek.

Najczęściej występującą grupą wiekową jest grupa 18-24. Wraz ze wzrostem wartości wieku spada liczebność klas. Skutkuje to przechyleniem wieku badanych osób w stronę młodości.

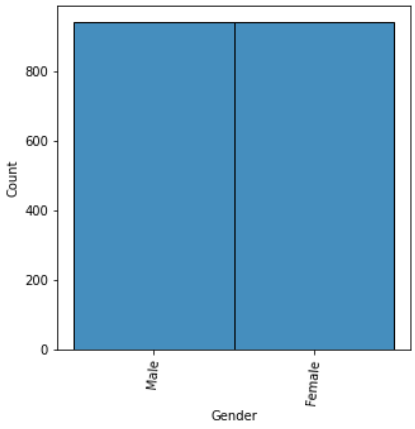
W przypadku płci zbiór jest podzielony niemalże idealnie na dwie równe części. Dzięki wartościom statystycznym zawartym w tabeli 1 można jednak zobaczyć, że badanych mężczyzn było nieznacznie więcej - można to odczytać, ponieważ skwantyfikowane wartości dla kobiet i mężczyzn mają taką samą wartość bezwzględną, z tym że wartość dla mężczyzn jest ujemna. Dla idealnego zbalansowania średnia powinna być równa 0, a obecna wartość to -0.000256, więc widoczny jest subtelnie większy udział płci męskiej.

Co do poziomu edukacji można odczytać, że największą część badanych osób stanowią osoby w trakcie studiów, bądź po studiach. Bazując na wiedzy o świecie można stwierdzić, że najczęściej to właśnie grupa ludzi w trakcie studiów może być podatna na różnego rodzaju używki, ze względu na tryb życia.

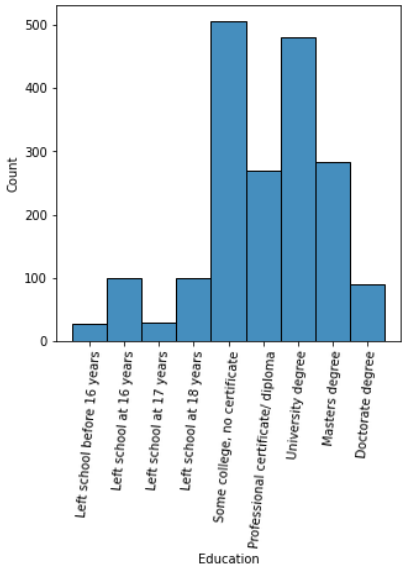
Najwięcej badanych mieszkało w Wielkiej Brytanii oraz Stanach Zjednoczonych. Pozostałe kraje stanowią małą część danych. Najbardziej liczna grupa co do pochodzenia etnicznego to grupa osób o białej skórze.



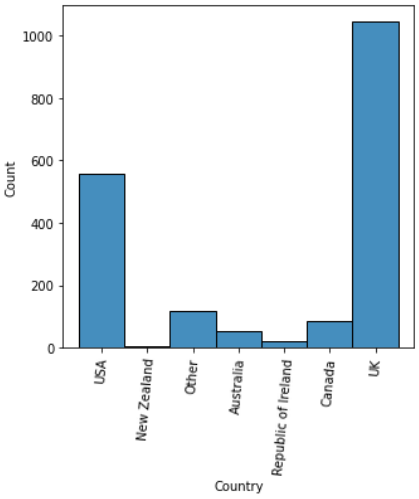
(a) Rozkład wieku uczestników



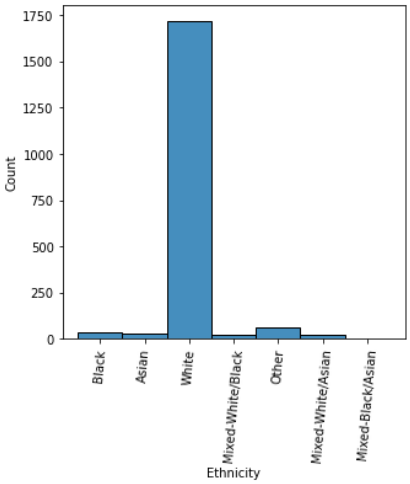
(b) Rozkład płci uczestników



(c) Rozkład poziomu edukacji uczestników

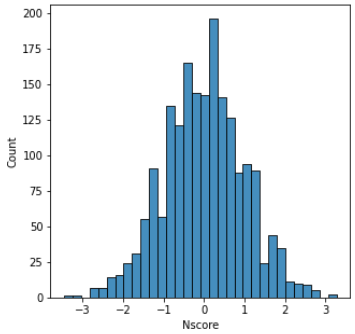


(d) Rozkład krajów pochodzenia uczestników

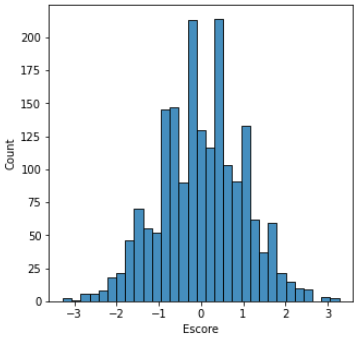


(e) Rozkład pochodzenia etnicznego uczestników

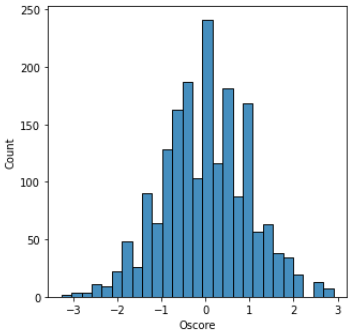
Rysunek 1: Rozkłady zmiennych dotyczących demografii badanych osób



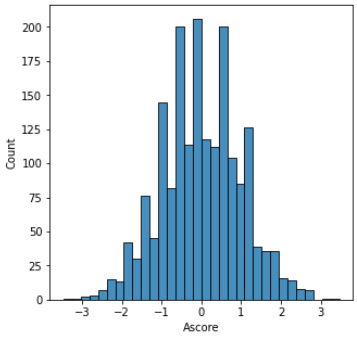
(a) Rozkład Nscore uczestników



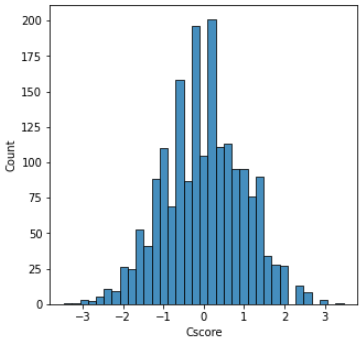
(b) Rozkład Escore uczestników



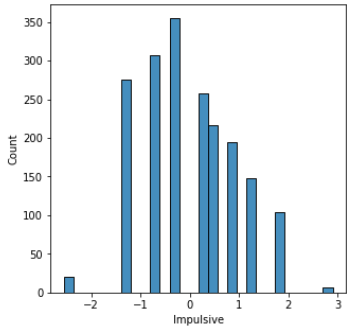
(c) Rozkład Oscore uczestników



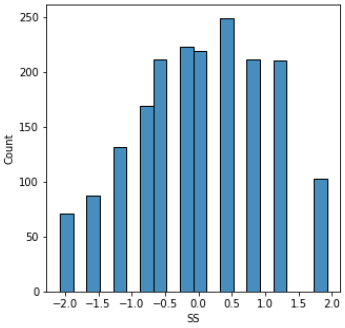
(d) Rozkład Ascore uczestników



(e) Rozkład Cscore uczestników



(f) Rozkład impulsywności uczestników



(g) Rozkład poszukiwania doznań (SS) uczestników

Rysunek 2: Rozkłady zmiennych dotyczących osobowości badanych osób

Cechy związane z osobowością, a dokładniej ich rozkłady w rozważanym zbiorze danych przedstawione zostały na rysunku 2. Jak można zobaczyć, przedstawione histogramy przypominają kształtem rozkłady normalne. Na bazie danych zawartych w tabeli 1 można określić, że te widoczne rozkłady normalne mają średnią około 0 oraz odchylenie standardowe około 1. Oczywiście wyniki te prezentują się w ten sposób ze względu na kwantyzację danych dokonaną przez autorów zbioru.

Na rysunku 3 zaprezentowane zostały rozkłady częstości stosowania poszczególnych używek. Jak widać, większość z nich jest mocno przechylona w jedną ze stron - albo większość badanych nigdy nie używała danej substancji, albo większość badanych używała jej w ciągu danego dnia. Najbardziej prawdopodobnym czynnikiem mającym wpływ na te zależności może być legalność poszczególnych używek. Używki takie jak alkohol, kofeina, czekolada czy nikotyna wśród ankietowanych najczęściej były używane w ciągu ostatniego dnia. Podobnie było również w przypadku marihuany, z tą różnicą, że niemalże tyle samo ankietowanych nigdy jej nie używała.

Nikotyna oraz marihuana są charakterystycznymi używkami, ze względu na to, że jako jedyne mają stosunkowo zbalansowany rozkład klas.

Używki, które są nielegalne, bądź są powszechnie uznawane za „mocniejsze", bardziej niebezpieczne czy też silniej uzależniające najczęściej nigdy nie były używane przez ankietowanych, a im wyższa częstotliwość stosowania, tym mniej ludzi udzielało takiej odpowiedzi.

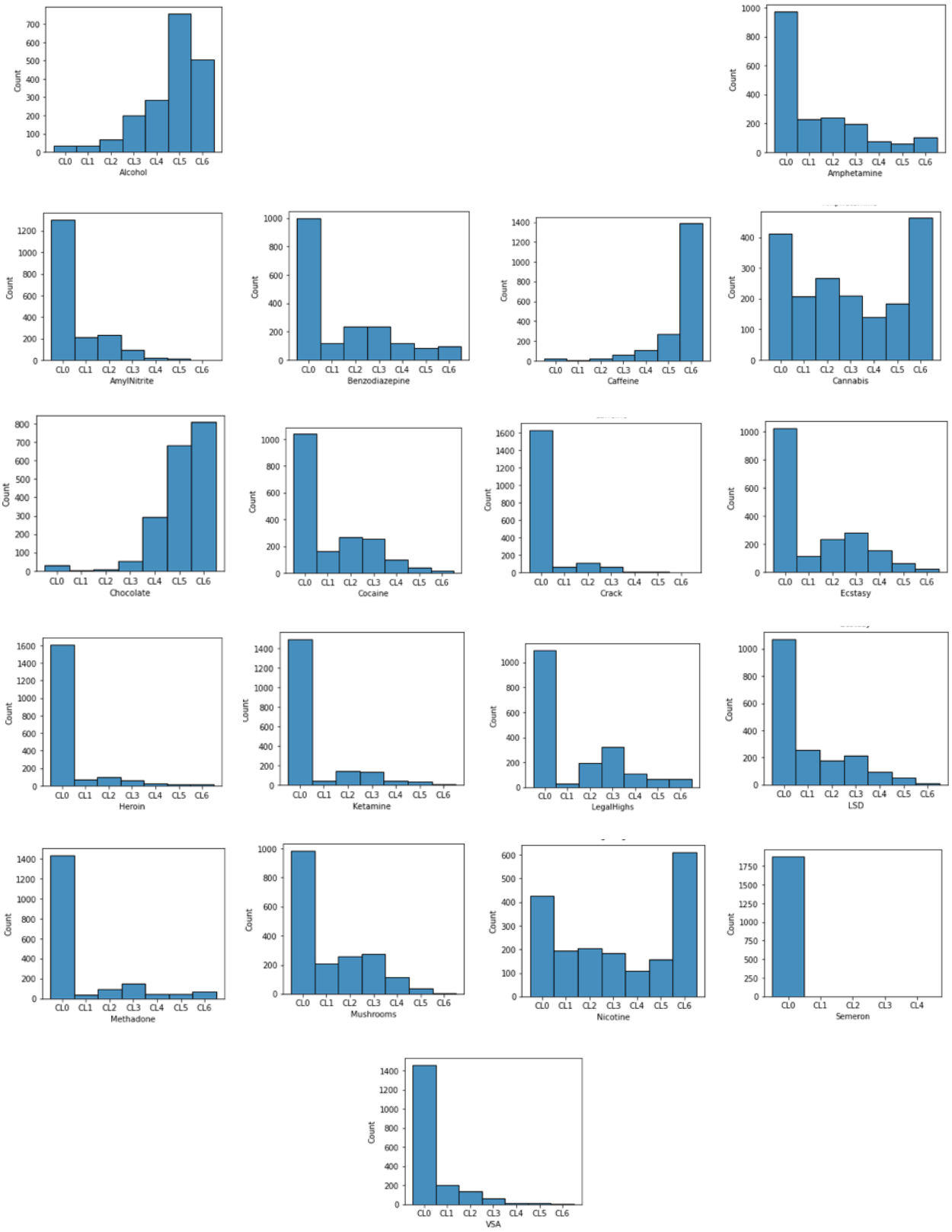
Warto zwrócić uwagę na używkę semeron - w tym przypadku wystąpiło najsilniejsze niezbalansowanie. Na 1885 przypadków zaledwie 8 należało do klasy innej niż CL0 (nigdy nie używano). W związku z powyższym używka ta zostanie pominięta w dalszych rozważaniach i analizach.

W związku z niezbalansowaniem każdej z klas podjęta została redukcja liczby klas. W jej wyniku powstały 3 klasy:

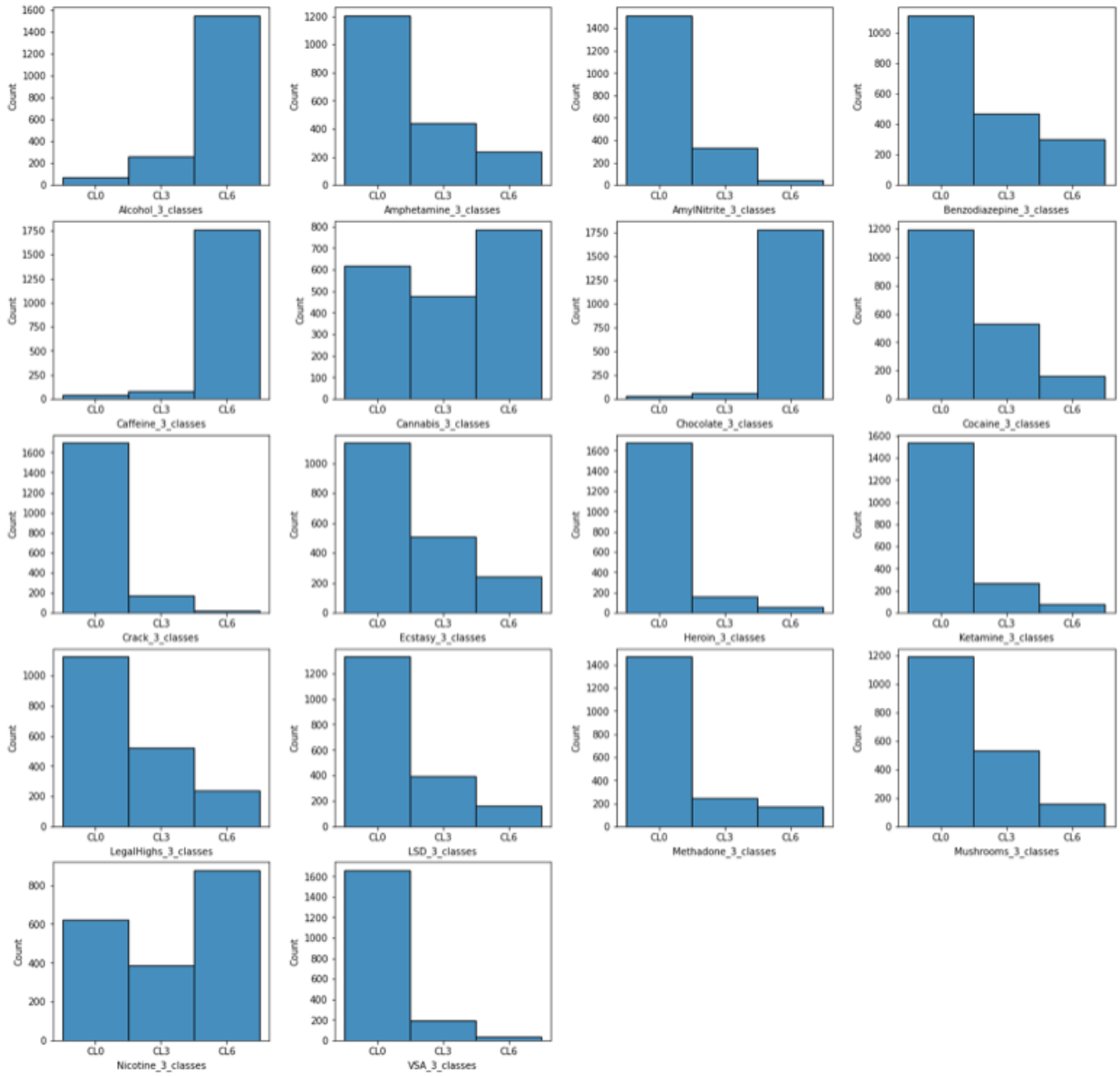
- CL0 - Nigdy nie używano - połączone CL0 (nigdy nie używano) oraz CL1 (używano ponad dekadę temu),
- CL3 - Używane w ostatniej dekadzie - połączone CL2 (używane w ostatniej dekadzie) i CL3 (używane w ciągu ostatniego roku),
- CL6 - używane w ostatnim miesiącu - połączone CL4 (używane w ostatnim miesiącu),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

Rozkłady dla zmodyfikowanego zbioru przedstawione zostały na rysunku 4. Charakterystyczne skrzywienia wykresów w stronę używania w ciągu ostatniego miesiąca oraz nie używania nigdy zostały zachowane tak jak w zbiorze bez zredukowanej liczby klas. Co jednak jest istotne nikotyna oraz marihuana stały się bliższe bycia zbalansowanymi. Można pokusić się o hipotezę, że te dwie używki prezentują podobne tendencje ze względu na sposób ich przyjmowania - każda z nich najczęściej używana jest w formie czegoś do palenia. Te dwie używki również są stosunkowo popularne u młodych ludzi, więc zważając na to, że znaczna część odpowiedzi została udzielona przez osoby poniżej 35 roku, nie powinno dziwić to, że tak prezentują się rozkłady.

Bazując na powyższych obserwacjach jako cel zadania obrana została klasyfikacja spożycia nikotyny oraz marihuany. Głównym powodem było stosunkowe zbalansowanie klas w porównaniu do innych używek oraz popularność obu z nich.



Rysunek 3: Rozkłady zmiennych dotyczących częstotliwości stosowania używek przez badane osoby

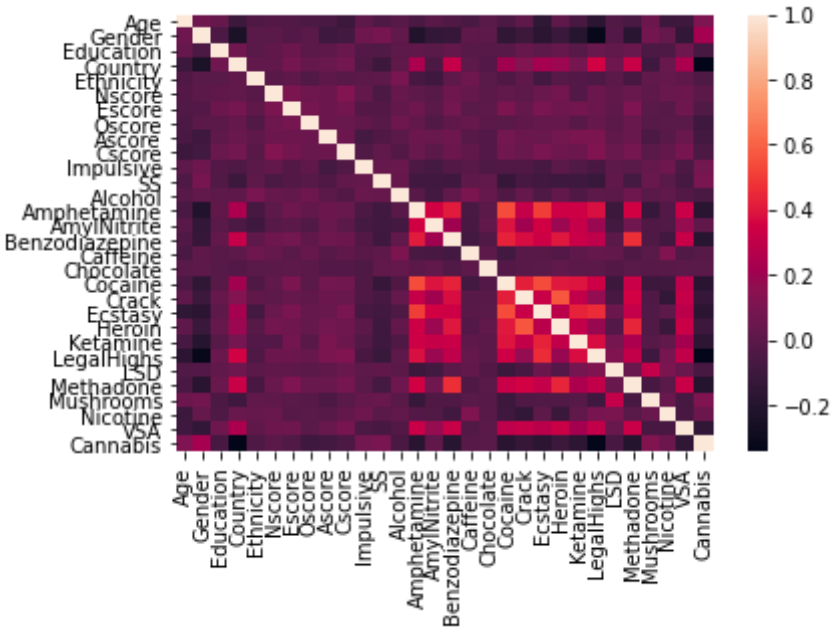


Rysunek 4: Rozkłady zmiennych dotyczących częstotliwości stosowania używek przez badane osoby ze zredukowaną liczbą klas

2.3 Zależności między parami zmiennych

W celu zbadania zależności pomiędzy wszystkimi parami zmiennych obliczona została korelacja Spearmana. Mierzy ona dowolną monotoniczną zależność między zmiennymi - także nieliniową. Wyniki zaprezentowane zostały na rysunku 5.

Jak można zauważyć najsilniejsza korelacja występuje pomiędzy niektórymi używkami. Po wgłębieniu się dotyczy to używek, które uznawane są za „mocniejsze", są nielegalne i silnie uzależniające. Może to oznaczać, że jeśli ktoś używa jednej z nich, to okazuje się, że używa też innych. Korelacja ta nie występuje w przypadku najpopularniejszych używek, czyli: alkoholu, kofeiny, czekolady, LSD, nikotyny oraz marihuany. Widoczne również są pewne zależności niektórych używek od aktualnego kraju zamieszkania. W pozostałych przypadkach nie ma widocznych korelacji między parami zmiennych.



Rysunek 5: Korelacja Spearmana dla wszystkich par zmiennych

W związku z tym, że docelowe zadanie realizowane w projekcie opierać będzie się na analizie używania nikotyny i marihuany sprawdzone zostały dokładne zależności tych wartości i każdej innej cechy. W przypadku stosowania innych używek do tej analizy pozostawiona została siedmiostopniowa skala (wartości opisane w podrozdziale 2.1).

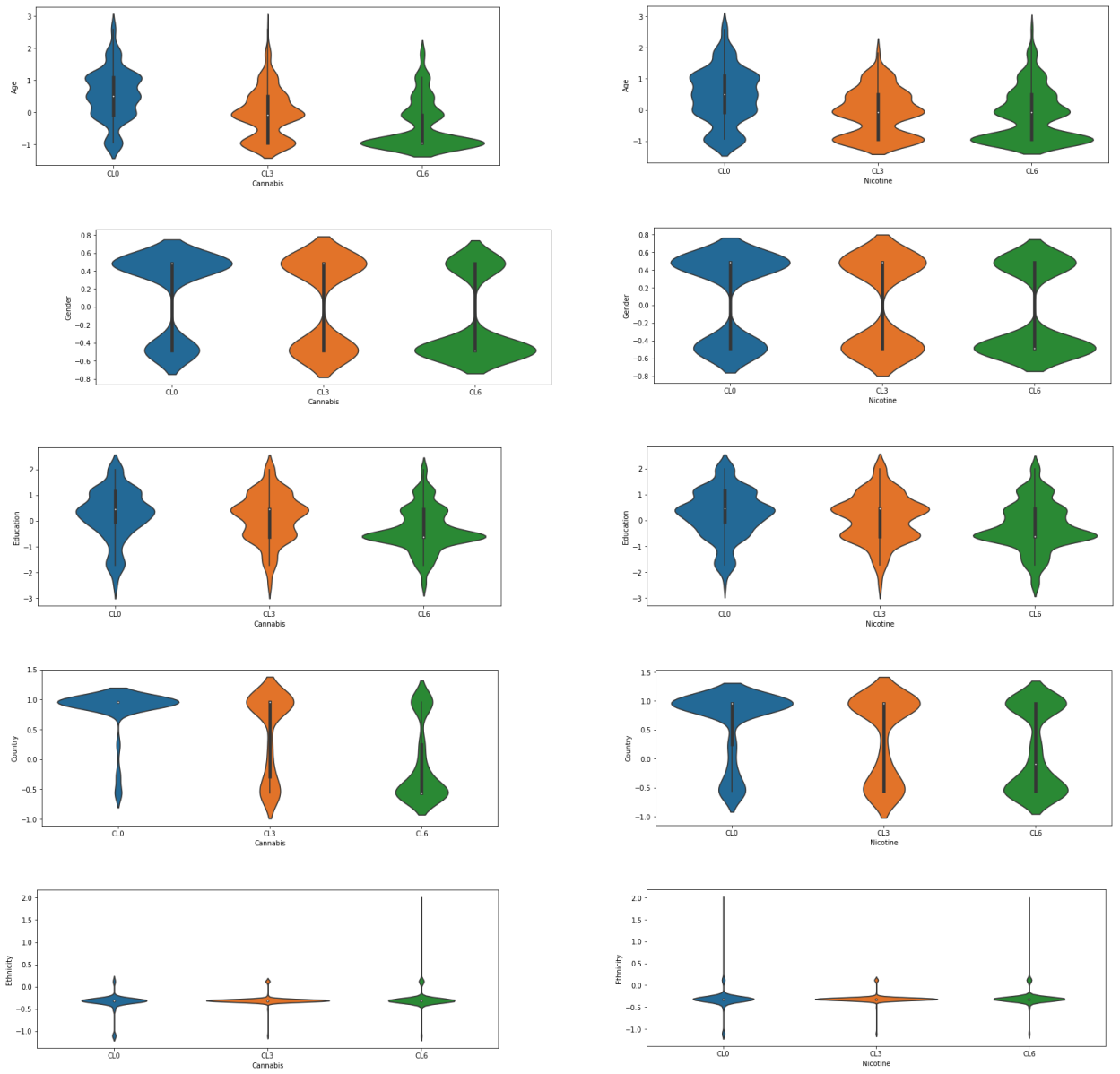
Poniżej zaprezentowane zostały wykresy skrzypcowe (ang. *violin plot*), które mówią o zależnościach pomiędzy parami zmiennych. Pierwszą rozważaną grupą, przedstawioną na rysunku 6 będą cechy demograficzne w zestawieniu z używaniem marihuany i nikotyny.

Jak widać prezentowane zależności są bardzo zbliżone. Potwierdza się wcześniej postawiona hipoteza - częściej po marihuanę i nikotynę sięgają najmłodsi ludzie. Widoczna również jest odwrotna zależność względem płci - mężczyźni (w obu przypadkach) częściej deklarują stosowanie tych używek w nieodległym czasie. Z kolei kobiety częściej odpowiadają, że nigdy nie stosowały żadnej z nich.

W przypadku zależności poziomu edukacji od częstego stosowania marihuany widoczny jest skok dla grupy, która uczęszcza na jakąś uczelnię, ale jeszcze jej nie ukończyła. Może to być również powiązane z wiekiem i trybem życia studentów, co już było wcześniej wspomniane. Podobne zależności zachodzą również dla nikotyny. Ciekawym jest też to, że zarówno dla marihuany jak i dla nikotyny dla klasy „używane w ostatniej dekadzie" (CL3) widoczne jest zwiększenie liczby odpowiedzi dla poziomu edukacji „Dyplom uniwersytetu" - może to świadczyć o tym, że osoby te próbowały używek w trakcie studiów, lecz po studiach ograniczyli ich spożycie.

Wpływ kraju zamieszkania na stosowanie używek może sprowadzać się głównie do legalności danej substancji w danym kraju. Więcej mieszkańców USA odpowiada, że stosowało marihuanę w przeciągu minionego miesiąca, a w przypadku Wielkiej Brytanii jest ich mniej. Dla nikotyny zależności te wyglądają podobnie, jednak wnioski o legalności nie są aplikowalne, ponieważ nikotyna jest legalna.

Pochodzenie etniczne nie niesie za sobą szczególnie istotnej informacji, zapewne ze względu na silne niezbalansowanie - dominującą wartością jest „mieszany afroamerykańsko-azjatycki".



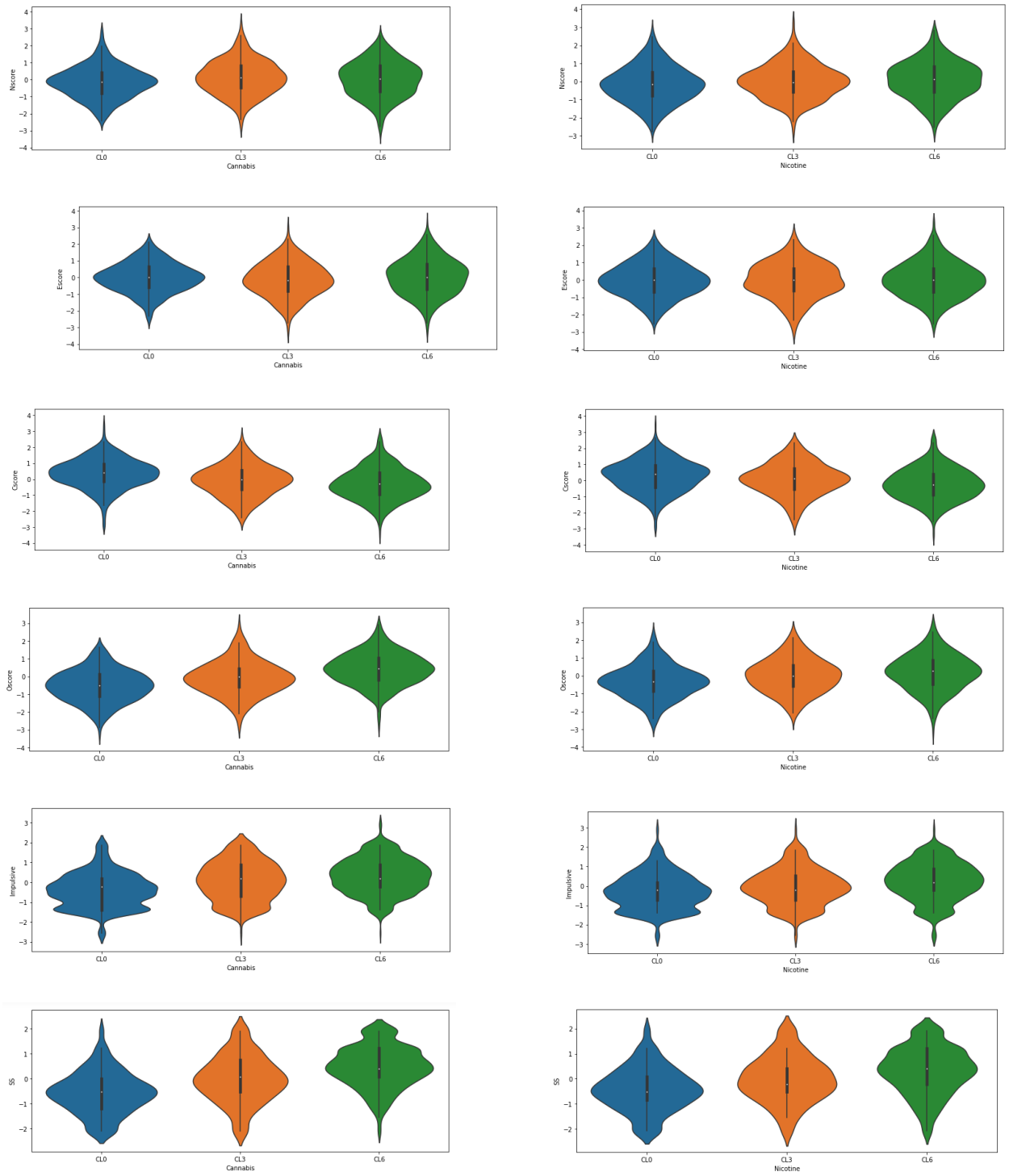
Rysunek 6: Wykresy skrzypcowe dla cech demograficznych i stosowania marihuany oraz nikotyny

Na rysunku 7 przedstawione zostały zależności cech dotyczących osobowości od stosowania marihuany i nikotyny. Wyglądają one niemalże identycznie dla obu używek. Ciekawe zależności widać w szczególności dla wartości Cscore, Oscore, impulsywności oraz poszukiwania doznań (SS).

Wartość średnia Cscore obniża się wraz z częstszym korzystaniem z tych używek. Należy zaznaczyć, że wartość Cscore odpowiada za sumienność. Osoby mniej konsekwentne mają często problemy z zaprzestaniem korzystania z używek, więc uzyskane wyniki wydają się być logiczne.

Z kolei w przypadku wartości Oscore występuje sytuacja odwrotna - im częściej stosowana jest dana używka, tym wyższy okazuje się być Oscore. Wartość ta odpowiada za otwartość na nowe doświadczenia, więc w tym przypadku rezultaty również można uznać za rozsądne.

Podobne zależności występują również dla impulsywności i poszukiwania nowych doznań - ludzie spokojni, nieszukający nowych doznań raczej stronią od używek.

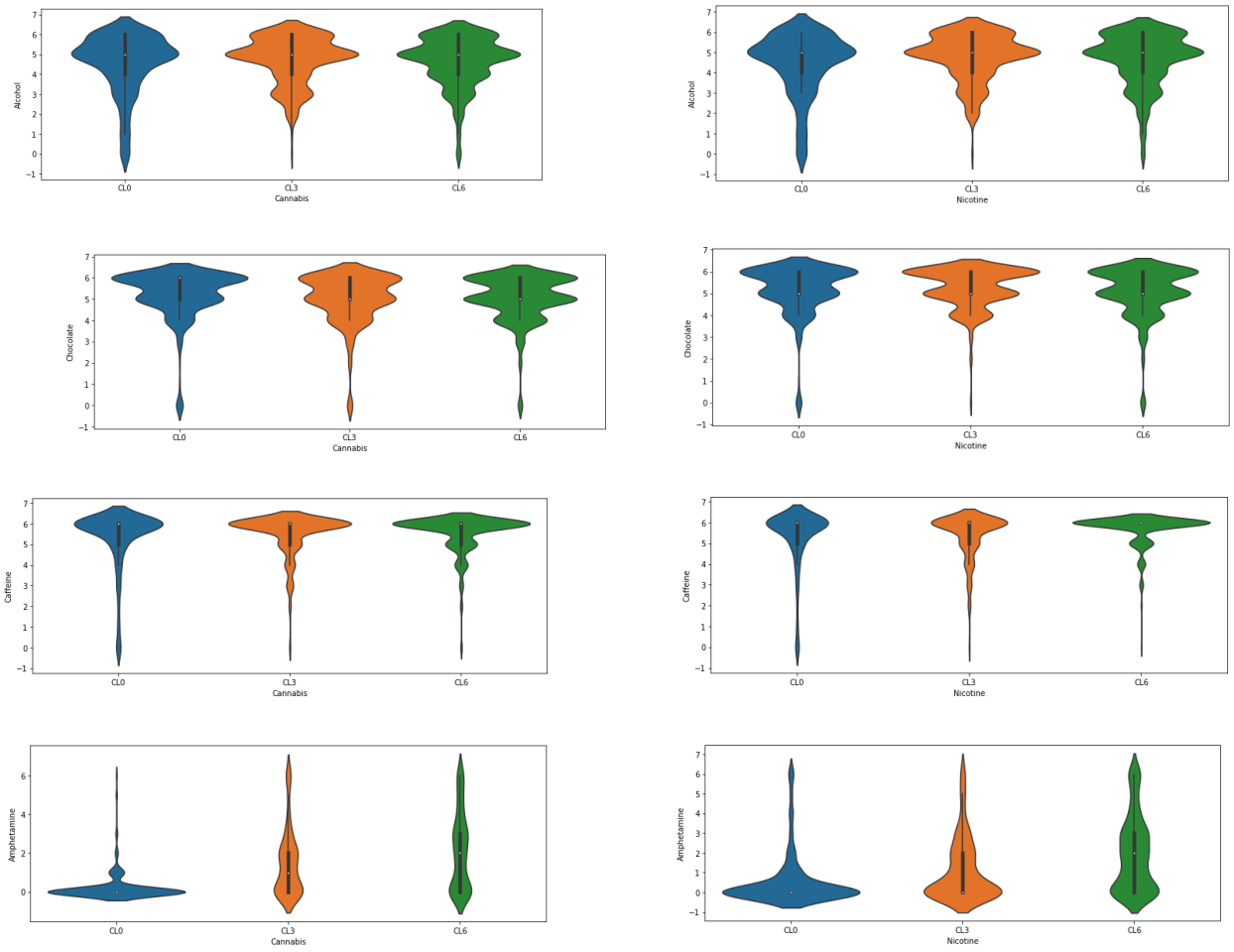


Rysunek 7: Wykresy skrzypcowe dla cech dotyczących osobowości i stosowania marihuany oraz nikotyny

Zależność stosowania marihuany i nikotyny od innych używek kolejny raz wygląda podobnie dla każdej z rozważanych używek. Można zauważyć, że używki, które są legalne (alkohol, czekolada oraz kofeina) są używane często, a stosowanie marihuany czy nikotyny nie ma większego znaczenia.

Przykład reprezentujący narkotyki „mocne”, nielegalne to amfetamina. Dla wszystkich środków zależności

były analogiczne - osoby rzadziej stosujące marihuanę czy nikotynę rzadziej sięgały po inne używki.



Rysunek 8: Wykresy skrzypcowe dla cech dotyczących stosowania alkoholu, czekolady, kofeiny oraz amfetaminy i stosowania marihuany oraz nikotyny

2.4 Przetwarzanie danych

W związku z tym, że zbiór danych był oryginalnie skwantyfikowany w ramach przetwarzania wykonana została jedynie standaryzacja. Dokonano jej przy użyciu narzędzia `StandardScales()` z biblioteki `sklearn`.

Dane zostały również podzielone na zbiór treningowy (70%) oraz testowy (30%). Podziału dokonano przy użyciu metody `train_test_split`, która również pochodzi z biblioteki *sklearn*. Oczywiści podział ten był stratyfikowany.

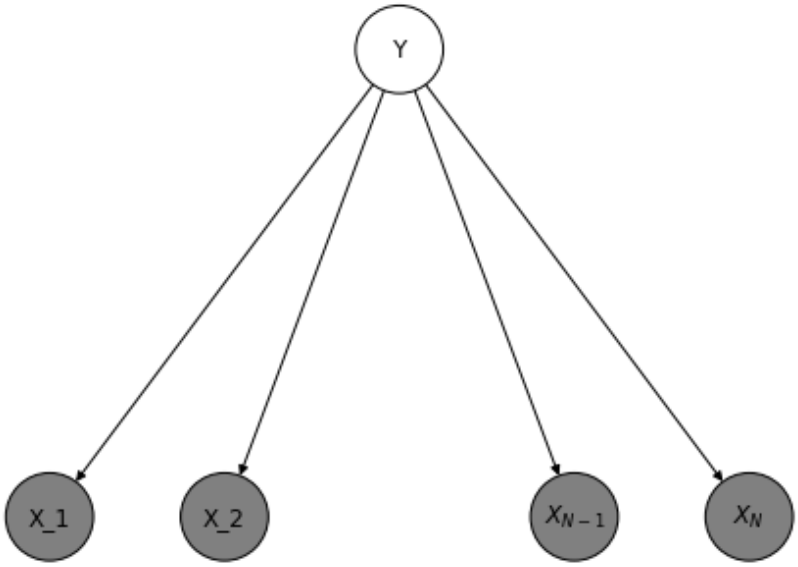
Przetworzone dane, wraz z przydzielonym typem użytkowania (zbiór treningowy lub zbiór testowy) zostały zapisane do dwóch plików: dla klasyfikacji spożycia marihuany do pliku `cannabis_preprocessed.csv`, a dla nikotyny analogicznie `nicotine_preprocessed.csv`. Pliki znajdują się w repozytorium w folderze `data`.

Ostatecznie zbiory zawierały 29 ustandaryzowanych cech - demograficznych, dotyczących osobowości oraz dotyczących stosowania innych używek. Każdy punkt danych przyjął jedną z trzech klas - CL0 (nigdy nie używano), CL3 (używano w ostatniej dekadzie) lub CL6 (używano w ostatnim miesiącu). Do zbioru treningowego trafiło 1319 rekordów, a do testowego 566.

3 Modele

3.1 Naiwny Bayes

Model Naiwnego Bayesa jest jednym z najprostszych modeli grafowych. Jego naiwność polega na założeniu, iż wszystkie obserwowane zmienne $X_1, X_2, \ldots, X_N$ są wzajemnie niezależne, a jedyną występującą zależnością ja ta występująca pomiędzy zmienną Y oraz zmiennymi obserwowanymi.



Rysunek 9: Przykład modelu grafowego opartego o reguły naiwnego Bayesa

Naszym zadaniem jest klasyfikacja, wiec zmienna Y będzie reprezentowała klasę, natomiast zmienne $X_1, X_2, \ldots, X_N$ będą naszymi cechami. Celem klasyfikatora jest znalezienie rzeczywistej klasy $\hat{y}$ dla danego zestawu cech. Uzyskiwany jest on poprzez maksymalizacje prawdopodobieństwa warunkowego klasy y_k pod warunkiem danych $x_1, x_2, \ldots, x_N$:

$$\hat{y} = \underset{k \in \{1,2,\ldots,K\}}{\operatorname{argmax}} P(y_k | x_1, x_2, \ldots, x_N)$$

Prawdopodobieństwo to, dzięki wykorzystaniu reguły Bayesa, możemy zapisać jako:

$$P(y_k | x_1, x_2, \ldots, x_N) = \frac{P(y_k) P(x_1, x_2, \ldots, x_N | y_k)}{P(x_1, x_2, \ldots, x_N)}$$

Licznik tego ułamka możemy zapisać jako prawdopodobieństwo łączne:

$$P(y_k) P(x_1, x_2, \ldots, x_N | y_k) = P(y_k, x_1, x_2, \ldots, x_N)$$

Dodatkowo możemy pominąć mianownik i zapisać, że prawdopodobieństwo warunkowe klasy y_k pod warunkiem danych $x_1, x_2, \ldots, x_N$ jest proporcjonalne do prawdopodobieństwa łącznego klasy oraz danych:

$$P(y_k | x_1, x_2, \ldots, x_N) \propto P(y_k, x_1, x_2, \ldots, x_N)$$

Korzystając wielokrotnie z reguły łańcuchowej możemy dokonać faktoryzacji prawdopodobieństwa łącznego:

$$\begin{aligned} P(y_k, x_1, x_2, \ldots, x_N) &= P(x_1, x_2, \ldots, x_N, y_k) \\ &= P(x_1|x_2, \ldots, x_N, y_k)P(x_2, \ldots, x_N, y_k) \\ &\ldots \\ &= P(x_1|x_2, \ldots, x_N, y_k)P(x_2|\ldots, x_N, y_k)\ldots P(x_{N-1}|x_N, y_k)P(x_N|y_k)P(y_k) \end{aligned}$$

Założenia wynikające z naiwności modelu (zmienne $X_1, X_2, \ldots, X_N$ są wzajemnie niezależne) pozwalają nam na uproszczenie:

$$P(x_i|x_{i+1}, x_{i+2}, \ldots, x_N, y_k) = P(x_i|y_k)$$

Podstawiając powyższe przekształcenie do wzoru wyżej otrzymujemy:

$$P(y_k|x_1, x_2, \ldots, x_N) \propto P(y_k, x_1, x_2, \ldots, x_N) = P(y_k)P(x_1|y_k)P(x_2|y_k)\ldots P(x_N|y_k) = P(y_k)\prod_{i=1}^{N} P(x_i|y_k)$$

$$P(y_k|x_1, x_2, \ldots, x_N) \propto P(y_k)\prod_{i=1}^{N} P(x_i|y_k)$$

Co ostatecznie daje nam:

$$\hat{y} = \underset{k\in\{1,2,\ldots,K\}}{\operatorname{argmax}} P(y_k)\prod_{i=1}^{N} P(x_i|y_k)$$

3.2 Gaussian Mixture Models

Modele mikstury rozkładów normalnych (ang. Gaussian Mixture Models, GMMs) to jeden z Latent Variable Models. Zakłada się w nich, że obserwowane dane są generowane z mikstury skończonej liczby rozkładów normalnych o nieznanych parametrach.

W przypadku modelu mikstur rozkładów normalnych mówi się o komponentach - zmiennych losowych o rozkładzie normalnym, z których każdy posiada:

- wektor średnich $\boldsymbol{\mu}$ określający środek rozkładu,
- macierz kowariancji $\boldsymbol{\Sigma}$ określający jego szerokość,
- mixing coefficient π mówiący o prawdopodobieństwie danego komponentu (ich suma musi być równa 1).

Wówczas, dla wektora losowego $\boldsymbol{x}$ funkcja gęstości GMM zdefiniowana jest w następujący sposób:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Stosując GMM do soft clustering można obliczyć prawdopodobieństwo posterior, że dany n-ty punkt należy do k-tego klastra - **responsibility**.

$$\gamma(z_{nk}) := p(z_n = k \mid x_n, \boldsymbol{\theta}) = \frac{p(z_i = k \mid \boldsymbol{\theta})\, p(x_n \mid z_n = k, \boldsymbol{\theta})}{\sum_{k'=1}^{K} p(z_n = k' \mid \boldsymbol{\theta})\, p(x_n \mid z_n = k', \boldsymbol{\theta})} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Stosując podejście probabilistyczne do obliczenia parametrów mikstur, okazauje się, że równania te nie mają jawnej postaci ($\gamma(z_{nk})$ zależy od pozostałych parametrów w złożony sposób i odwrotnie). Dlatego stosuje się podejście iteracyjne. W tym przypadku algorytm Expectation-Maximization (EM).

Algorytm EM dla GMM będzie postępował według następujących kroków:

1. Zainicjalizuj parametry $\boldsymbol{\theta}$ ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}$)
2. Oblicz responsibilities $\gamma(z_{nk})$ dla aktualnych wartości parametrów (expectation step)
3. Oblicz i zaktualizuj parametry ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\pi}$) wykorzystując wartości responsibilities (maximization step).

Algorytm jest zatrzymywany po osiągnięciu określonej liczby kroków, gdy zmiana log-likelihood jest poniżej progu lub gdy zmiana parametrów jest poniżej progu.

W ogólności modele mikstur rozkładów normalnych służą do klasteryzacji danych.

4 Eksperymenty

4.1 Naiwny Bayes

W badaniach porównane zostały różne implementacje naiwnego klasyfikatora Bayesa:

- implementacje z biblioteki *sklearn*:
  - $GaussianNB$,
  - $MultinomialNB$,
  - $ComplementNB$,
  - $BernoulliNB$,
  - $CategoricalNB$,
- implementacja z wykorzystaniem biblioteki *pyro*,
- implementacja z wykorzystaniem biblioteki *pgmpy*.

Wykorzystane dane są dyskretne, ale podane są jako wartości $float$. Niestety w zbiorze testowym dla niektórych cech zdarzają się wartości, które nie wystąpiły w zbiorze treningowym. Z tego powodu, dla klasyfikatorów, które wymagają danych dyskretnych ($CategoricalNB$ oraz implementacja wykorzystująca bibliotekę *pgmpy*), zastosowana została dyskretyzacja $KBins$ z biblioteki *sklearn*. Dodatkowo klasyfikatory $MultinomialNB$ oraz $ComplementNB$ nie przyjmują danych ujemnych, wiec w ich przypadku zastosowany został $MinMaxScaler$ z biblioteki *sklearn*. Pozostałe klasyfikatory otrzymały oryginalne dane.

Badania

Klasyfikatory zostały przetestowane pod względem przyjmowanych danych - cechy opisujące danego człowieka albo cechy wzbogacone o częstotliwość spożywania innych używek.

W przypadku klasyfikatorów przyjmujących dane dyskretne zostały sprawdzone wyniki w zależności od wartości K dla dyskretyzatora $KBins$.

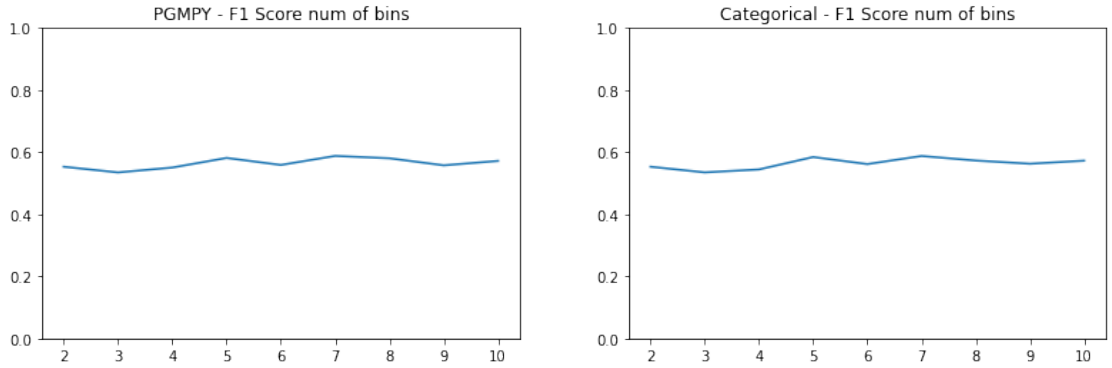
Dla klasyfikatora zaimplementowanego z wykorzystaniem biblioteki *pyro* została sprawdzone potrzebna liczba epok, aby zminimalizować funkcje straty.

Główne miary do oceny jakości klasyfikacji:

- F1 Score macro,
- ROC AUC macro OVR (One Versus Rest).

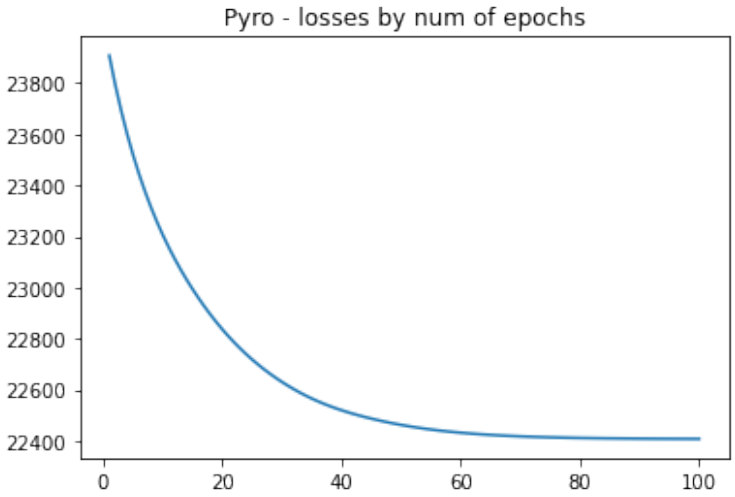
4.1.1 Wyniki dla zbioru opisującego spożycie marihuany

Cechy związane z człowiekiem

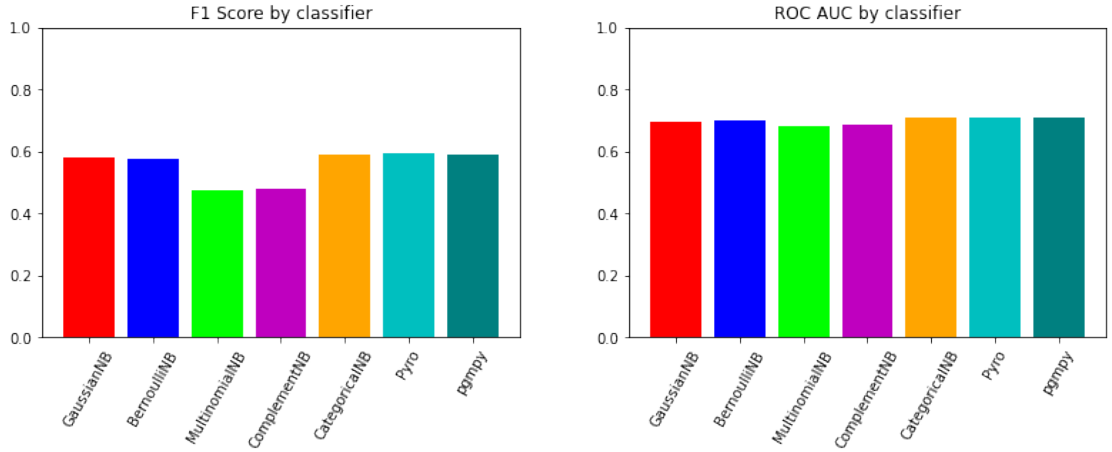


(a) Model *pgmpy* (b) CategoricalNB

Rysunek 10: Wartości miary F1 w zależności od ilości podziałów dla dyskretyzatora



Rysunek 11: Wartość funkcji straty w zależności od liczby epok

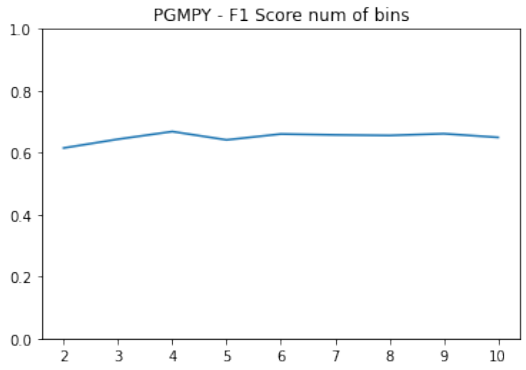


(a) F1 Score

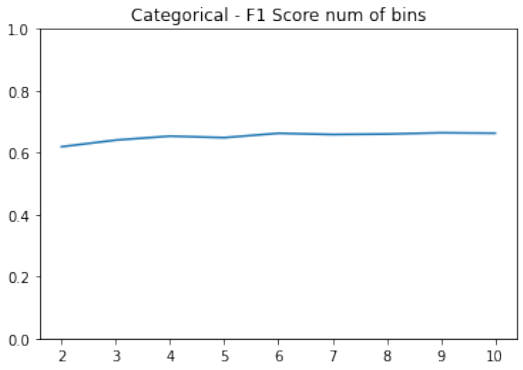
(b) ROC AUC

Rysunek 12: Wartości miary F1 oraz ROG AUC w zależności od użytego modelu

Wszystkie cechy

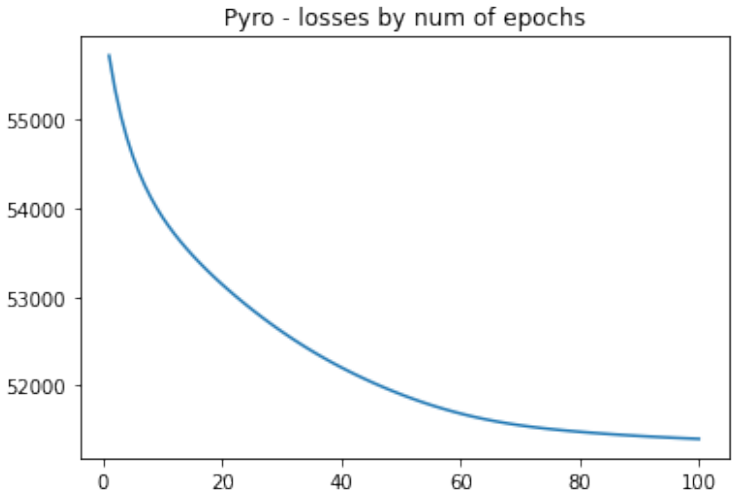


(a) Model *pgmpy*

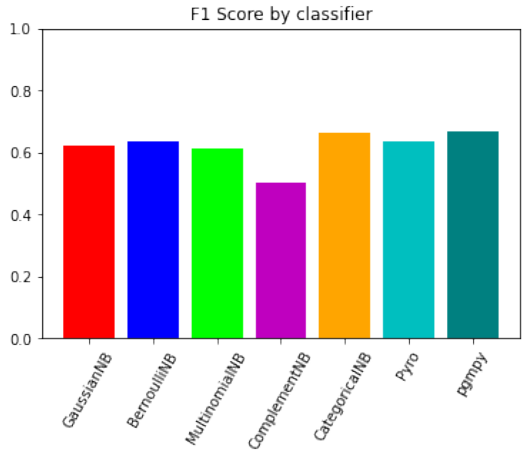


(b) CategoricalNB

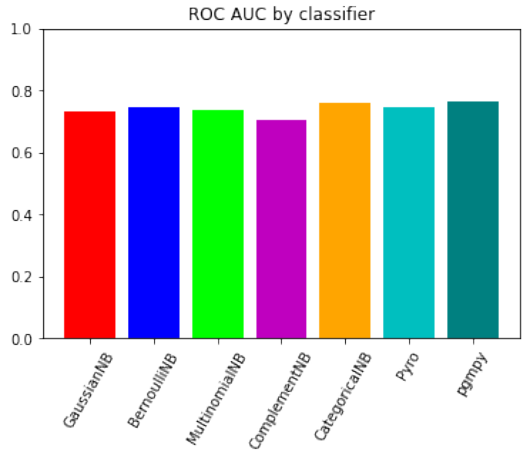
Rysunek 13: Wartości miary F1 w zależności od ilości podziałów dla dyskretyzatora



Rysunek 14: Wartość funkcji straty w zależności od liczby epok



(a) F1 Score



(b) ROC AUC

Rysunek 15: Wartości miary F1 oraz ROG AUC w zależności od użytego modelu

Wnioski

Można zauważyć, że wartość parametru K dla dyskretyzatoa $KBins$ nie miała dużego znaczenia. Najlepsze wyniki udało się uzyskać dla $K = 5$ dla cech związanych z człowiekiem oraz $k = 4$ dla wszystkich cech - różnica między innymi podziałami jest jednak znikoma.

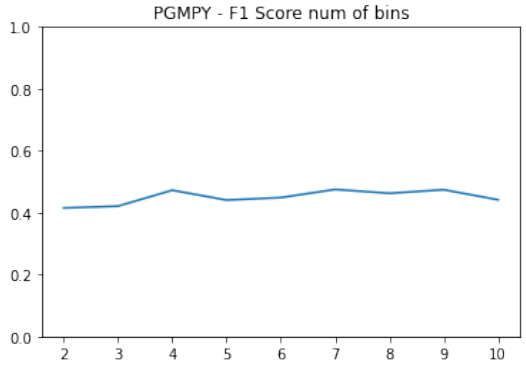
W przypadku klasyfikatora z biblioteki *pyro* możemy zauważyć, iż potrzebuje on większej liczby epok aby zbiec do minimum funkcji straty.

Patrząc na miary $F1$ oraz $ROCAUC$ można zauważyć, że prawie wszystkie klasyfikatory osiągają podobne wyniki. Wyjątkami są modele $MultimodalNB$ oraz $ComplementNB$, z czego ten pierwszy poradził sobie zdecydowanie lepiej w przypadku wykorzystania wszystkich cech. Warto również zaznaczyć, że klasyfikatory miały największy problem z klasą $CL3$, która jest najmniej reprezentowana w zbiorze. Dodatkowo zauważyć można, iż miara F1 lepiej pokazuje różnice w jakości klasyfikatorów.

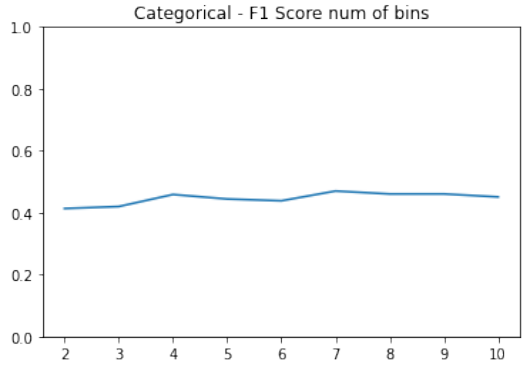
Porównując wykorzystane cechy łatwo dostrzec, iż wykorzystanie dodatkowych informacji o spożywaniu innych używek pozwoliło poprawić wyniki wszystkich klasyfikatorów.

4.1.2 Wyniki dla zbioru opisującego spożycie nikotyny

Cechy związane z człowiekiem

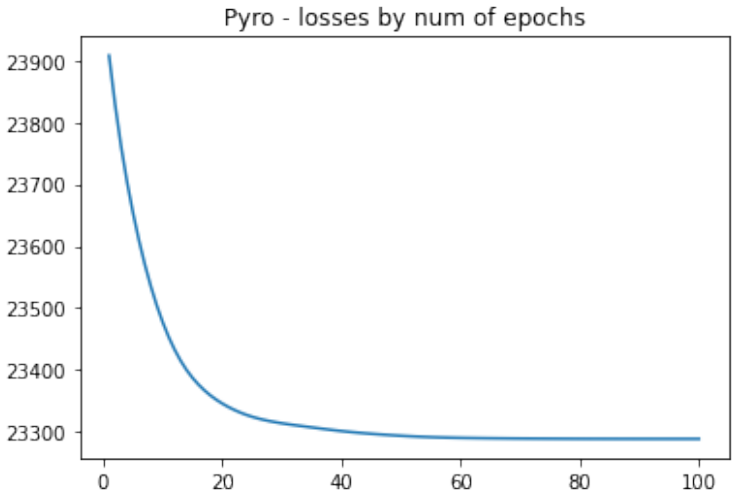


(a) Model *pgmpy*

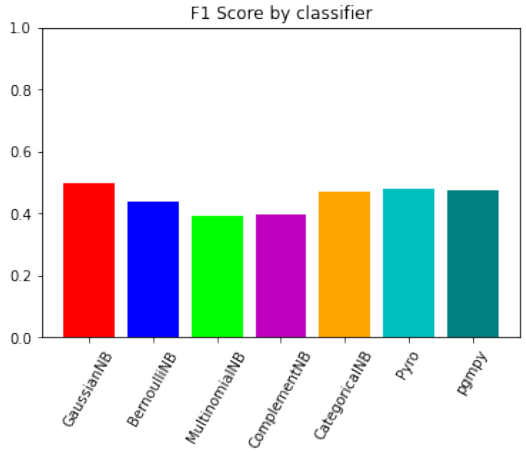


(b) CategoricalNB

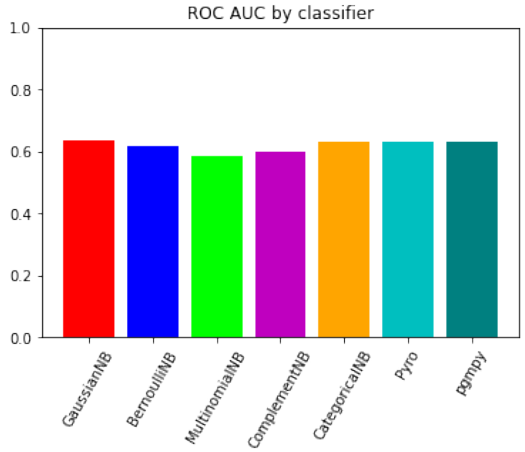
Rysunek 16: Wartości miary F1 w zależności od ilości podziałów dla dyskretyzatora



Rysunek 17: Wartość funkcji straty w zależności od liczby epok



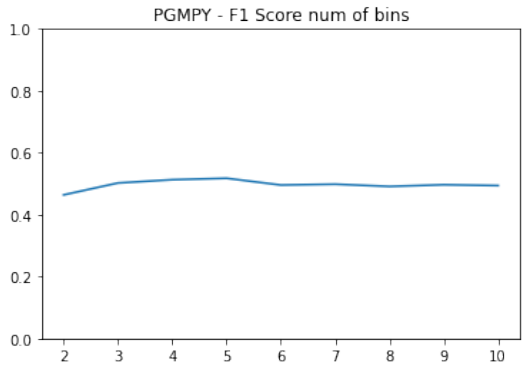
(a) F1 Score



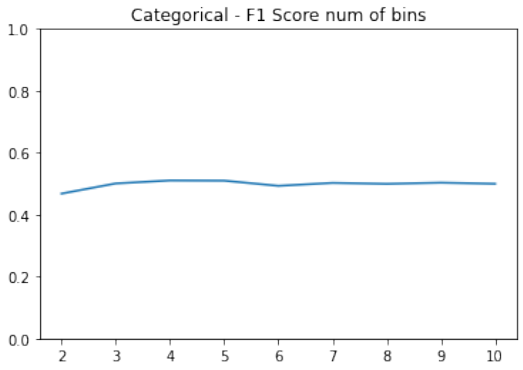
(b) ROC AUC

Rysunek 18: Wartości miary F1 oraz ROG AUC w zależności od użytego modelu

Wszystkie cechy

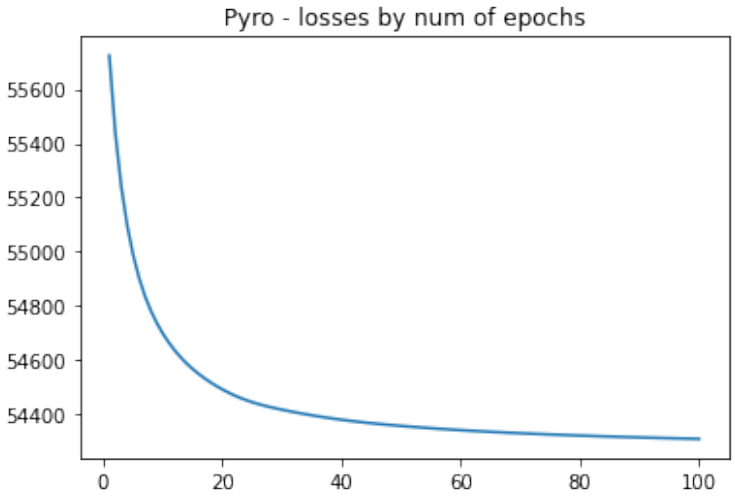


(a) Model *pgmpy*

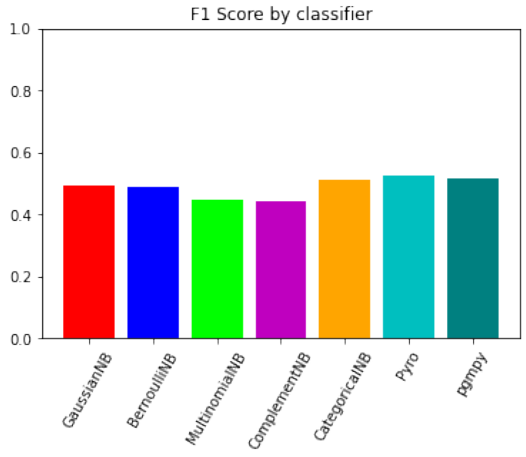


(b) CategoricalNB

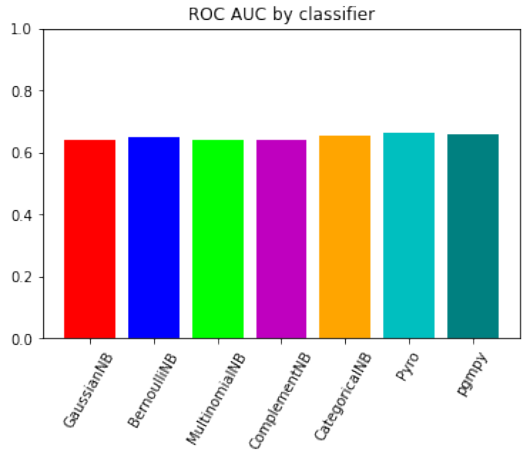
Rysunek 19: Wartości miary F1 w zależności od ilości podziałów dla dyskretyzatora



Rysunek 20: Wartość funkcji straty w zależności od liczby epok



(a) F1 Score



(b) ROC AUC

Rysunek 21: Wartości miary F1 oraz ROG AUC w zależności od użytego modelu

Wnioski

W przypadku drugiego zbioru znowu w oczy rzuca się,że wartość parametru K dla dyskretyzatora $KBins$ nie miała dużego znaczenia.

Widać jednak nieco inne wyniki w przypadku liczby potrzebnych epok dla klasyfikatora z biblioteki *pyro* - model ten zbiega zauważalnie szybciej niż w przypadku poprzedniego zbioru, a ponadto nie widać znaczącej różnicy jeśli chodzi o tą wartość w przypadku wykorzystania wszystkich cech albo tylko tych związanych z człowiekiem.

Patrząc na miary $F1$ oraz $ROCAUC$ znowu można zauważyć, że praktycznie wszystkie klasyfikatory osiągają podobne wyniki. Wraz z modelami $MultimodalNB$ i $ComplementNB$ nieco gorszą jakość wydaje się mieć klasyfikator $BernoulliNB$. Modele znowu miały największy problem z klasą $CL3$ - w tym przypadku jest ona jeszcze mniej reprezentowana niż w problemie dotyczącym spożycia marihuany.

Jeżeli chodzi o porównanie wykorzystanych cech, to tak samo jak w przypadku poprzedniego zbioru - dodanie dodatkowych cech nieco poprawiło wyniki. Zysk jest jednak zauważalnie niższy niż w poprzednim przypadku Porównując wykorzystane cechy łatwo dostrzec, iż wykorzystanie dodatkowych informacji o spożywaniu innych używek pozwoliło poprawić wyniki wszystkich klasyfikatorów.

4.2 Podsumowanie

Podsumowując, ciężko wybrać jest jeden najlepszy klasyfikator Naiwnego Bayesa. W przypadku badanych problemów najlepiej zadziałały modele $GaussianNB$, $CategoricalNB$ oraz te implementowane z wykorzystaniem bibliotek *pyro* czy *pgmpy*.

Jednoznacznie można powiedzieć coś odnośnie cech wykorzystanych jako wejście modelu. Mianowicie większa liczba cech pozwoliła na uzyskanie lepszych wyników w obu badanych przypadkach.

Warto również zauważyć że użyte zbiory były niezbalansowane, a modele miały największy problem z klasyfikacją klasy najsłabiej reprezentowanej. W tym przypadku do oceny jakości klasyfikacji bardzo dobrze sprawdziła się miara $F1macro$.

4.3 Modele mikstury rozkładów normalnych

W badaniach porównano dwie implementacje modeli mikstury rozkładów normalnych:

- pochodzącej z biblioteki *sklearn*,
- napisaną od podstaw (ang. from scratch)

4.3.1 Badania

Jako, że modele mikstur rozkładów normalnych służą do klasteryzacji, a nie klasyfikacji podstawowe metryki nie działają. Niemożliwa jest pewna identyfikacja klastra z klasą. Modele mikstur zwracają przynależność nowych próbek do klastrów, jednak w sposób dosłowny nie określają, który klaster odnosi się do której klasy. Dlatego w celach badawczych użyto miar AIC i BIC (opisane dokładniej w rozdziale 4.2.2) oraz wizualizacji.

W badaniach porównano hiperparametry:

- maksymalną liczbę iteracji: [10, 30, 50, 100],
- liczbę komponentów: [2, 3, 4, 5, 6, 7, 8, 9, 10]

Pod uwagę brano 12 cech, które opisują człowieka (szerzej opisane w rozdziale 2.1). W celu wizualizacji oraz lepszego poglądu na rozkład danych użyto metody do reduckji wymiarowości **PCA** (ang. Principal component analysis). Klasteryzacji dokonano na dwóch zbiorach, które dotyczyły spożycia marihuany oraz nikotyny.

Dane zostały podzielone za zbiór treningowy i testowy w proporcji 30/70. Miary AIC oraz BIC obliczane były tylko na zbiorze treningowym, a wnioski z analizy wizualizacji wyciągano na podstawie zbioru testowego.

4.3.2 Użyte miary

Na początku do wyznaczenia najlepszych hiperparaetrów sprawdzone zostało, który z modeli osiąga największą wartość likelihooda. W literaturze najczęściej dla tego typu problemów wykorzystuje się **Akaike Information Criterion (AIC)**

$$AIC = \ln\left(\hat{L}\right) - M$$

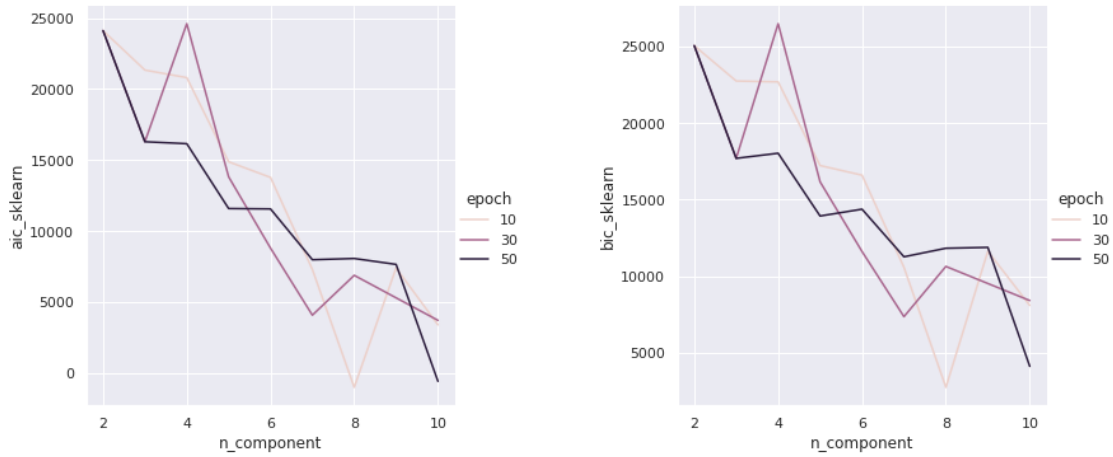
oraz **Bayesian Information Criterion (BIC)**. Wybór modelu oparty jest na maksymalnej wartości kryterium.

$$BIC = \ln\left(\hat{L}\right) - \frac{1}{2}M \ln N$$

gdzie M to liczba estymowanych parametrów w modelu, N to liczba punktów w danych, a $\ln\left(\hat{L}\right)$ to najwyższa wartość log-likelihood. Porównując AIC i BIC widać, że BIC bardziej personalizuje złożoność modelu.

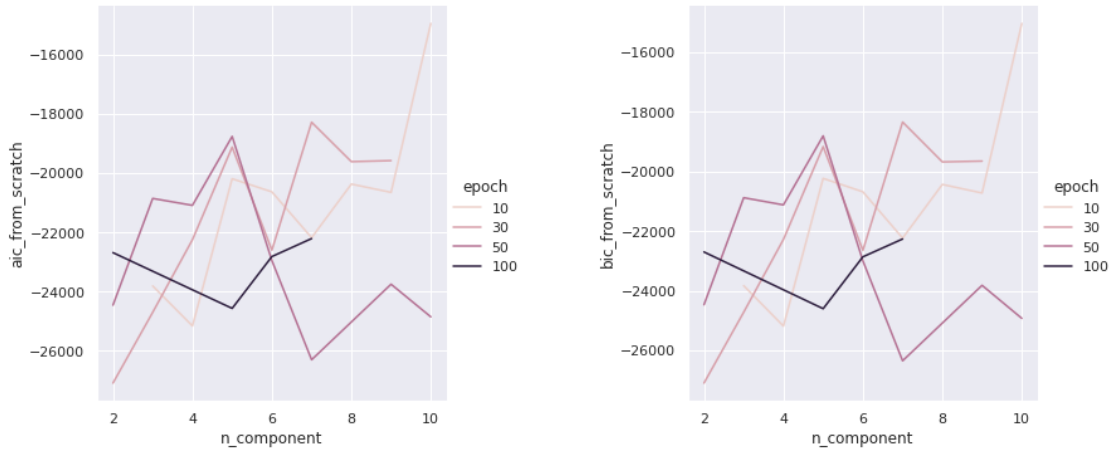
4.3.3 Wyniki dla zbioru opisującego spożycie marihuany

Pierwszym krokiem było porównanie miar AIC oraz BIC w zależności od różnych wartości hiperparametrów. Wyniki przedstawione zostały na Rysunku 22 dla modelu z biblioteki *sklearn*, oraz na Rysunku 23 dla własnej implementacji modelu. Podsumowując, najlepsze wyniki dla modeli przedstawia Tablica 2.



(a) Wartości miary AIC (b) Wartości miary BIC

Rysunek 22: Miary AIC oraz BIC w zależności od liczby iteracji dla modelu z biblioteki *sklearn*



(a) Wartości miary AIC (b) Wartości miary BIC

Rysunek 23: Miary AIC oraz BIC w zależności od liczby iteracji dla własnej implementacji modelu

Model	# iteracji	# komponentów	AIC	BIC
sklearn	10	2	24119.05	25057.47
sklearn	30	2	24119.05	25057.47
sklearn	50	2	24119.05	25057.47
scratch	10	10	-14951.43	-15029.20
scratch	30	7	-18284.49	-18338.93
scratch	50	5	-18764.50	-18803.38

Tablica 2: Najlepsze wartości hiperparametrów względem miar AIC i BIC dla obu modeli

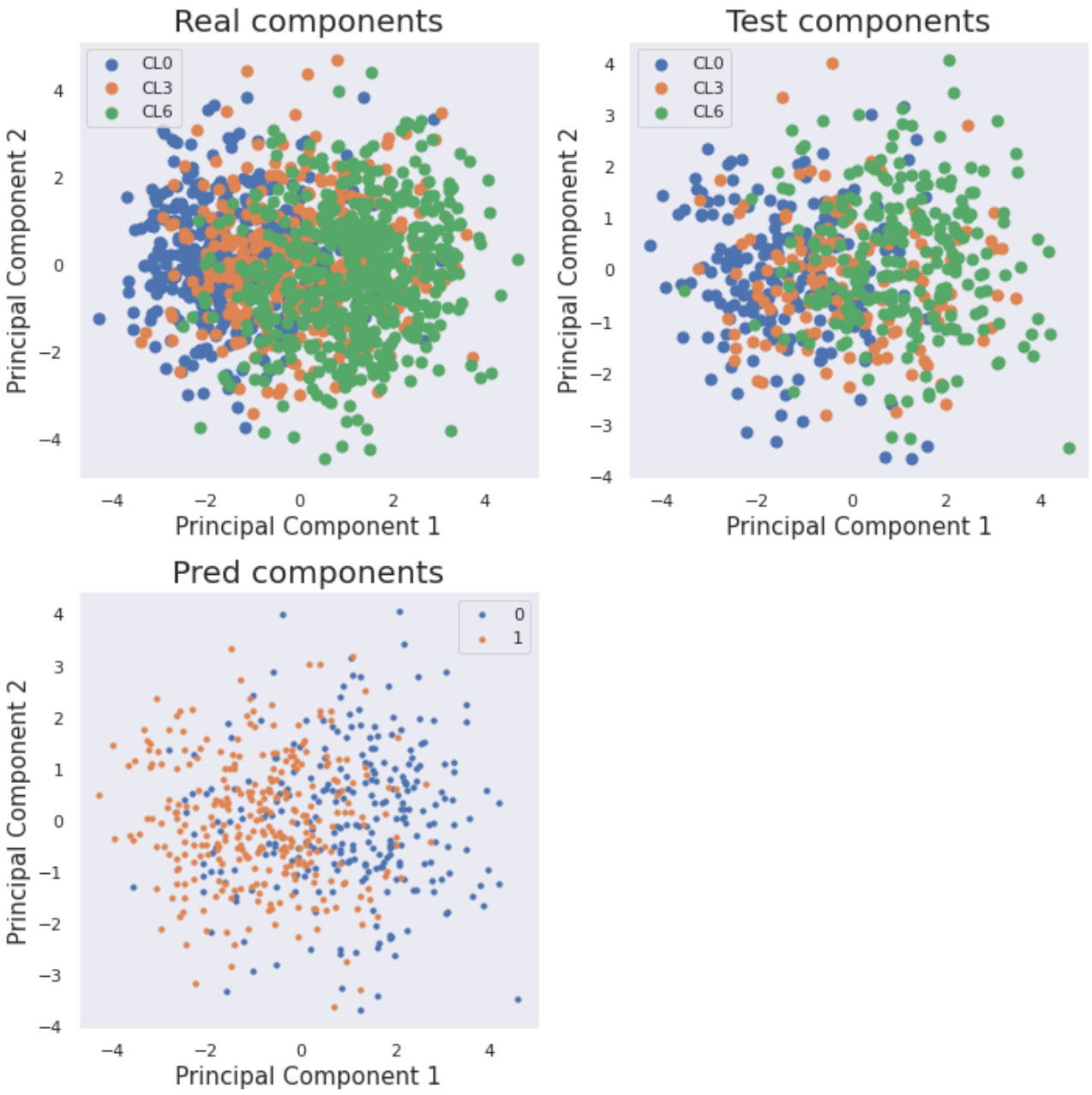
W przypadku modelu z *sklearna* wartości obu metryk są dodatnie, a w przypadku modelu własnego są ujemne. Może to świadczyć o tym, że model z wbudowanej biblioteki został lepiej skonstruowany.

Dodatkowo rozkłady najlepszych hiperparametrów w przypadku obu modeli znacząco się różnia. Dla modelu z *sklearna* wyniki AIC i BIC dla różnych wartości iteracji są takie same. Oznacza to, że 10 iteracji wystarczyło aby model osiągnął najlepszy wynik. Na wykresach na Rysunku 22 można zauważyć, że liczba iteracji nie wpływa tak znacząco na wyniki jak w przypadku modelu własnej implementacji, które przedstawione są na Rysunku 23. Dla różnych wartości iteracji, przy tych samych wartościach liczby komponentów wyniki są różne. Może to oznaczać, że model jest mniej stabilny. Dodatkowo widać, że większa liczba iteracji niekoniecznie oznacza najlepsze wyniki. Jest wręcz odwrotnie, ponieważ dla 10 iteracji wyniki są najlepsze.

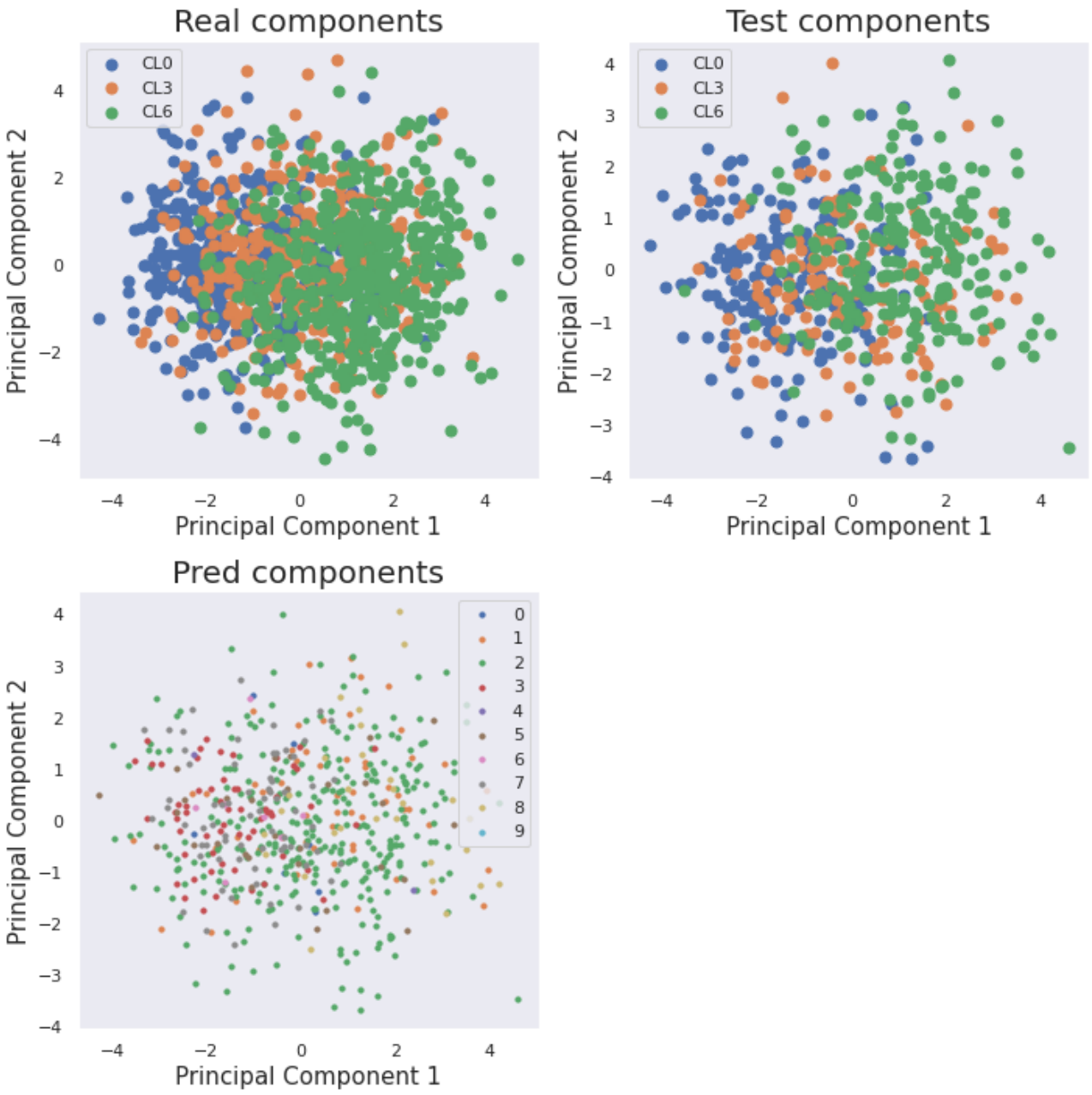
Na Rysunku 33 przedstawiona zostały wizualizacja dla najlepszych hiperparametrów modelu z biblioteki *sklearn*. Obie osie odnoszą się do komponentów powstałych dzięki analizie PCA. Pierwszy rysunek odnosi się do prawdziwych danych ze zbioru treningowego i ich podziału na trzy klasy. Na kolejnym zaznaczono prawidłowe wartości podziału danych w zbiorze testowym. Na ostatnim rysunku widać jakie przewidział wyniki. Zauważyć można, że model nauczył się generalizacji. Połączył dwie klasy CL0 oraz CL3 w jedną. Jeśli przyjrzeć się ich znaczeniu, oznacza to, że złączył przypadki w którym osoba nigdy nie spożywała używek z przypadkami gdy spożywała je w ostatniej dekadzie lub ostatnim roku. Rozróżnia je od przypadków gdy osoba spożywała używki w ostatnim miesiącu, tygodniu lub dniu.

Na Rysunku 25 przedstawiona została wizualizacja dla najlepszych hiperparametrów modelu własnej implementacji. Jak poprzednio, na dwóch pierwszych rysunkach zostały przedstawione kolejno: podział klastrów na zbiorze treningowym, prawidłowy podział na zbiorze testowym oraz przewidziany podział na zbiorze testowym. W tym przypadku najlepsze wartości miar AIC oraz BIC model uzyskał dla dziesięciu klastrów. Kolor zielony i pomarańczowy, czyli klastry o etykiecie 2 oraz 1 pojawiają się w każdym miejscu wykresu i ciężko znaleźć dla nich odpowiadające im etykiety. W przypadku reszty klas sytuacja jest lepsza i widać podział na trzy klasy. Klastry: 0, 7 odpowiadają etykiecie CL0, klastry 3, 5, 6 etykiecie CL3, a klastry 4, 8, 9 etykiecie CL6.

Podsumowując, model z wbudowanej biblioteki lepiej generalizuje dane, a ten własnoręcznie zaimplementowany lepiej podzielił zbiór na klasy. Warto również zaznaczyć, że oryginalnie dane posiadają 7 etykiet, co może sugerować, że własna implementacja w tym przypadku zadziałała lepiej.



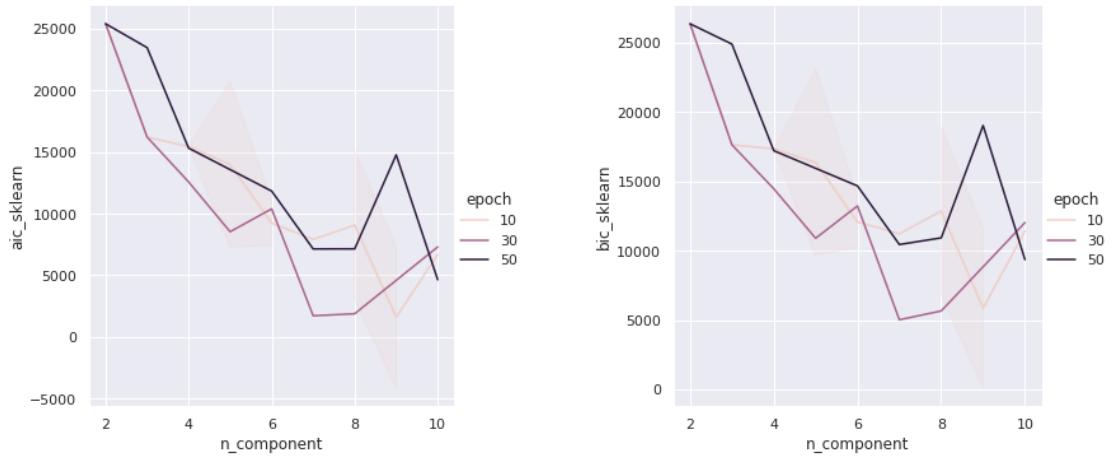
Rysunek 24: Wizualizacja danych dla najlepszych hiperparametrów modelu z biblioteki *sklearn*



Rysunek 25: Wizualizacja danych dla najlepszych hiperparametrów modelu własnej implementacji

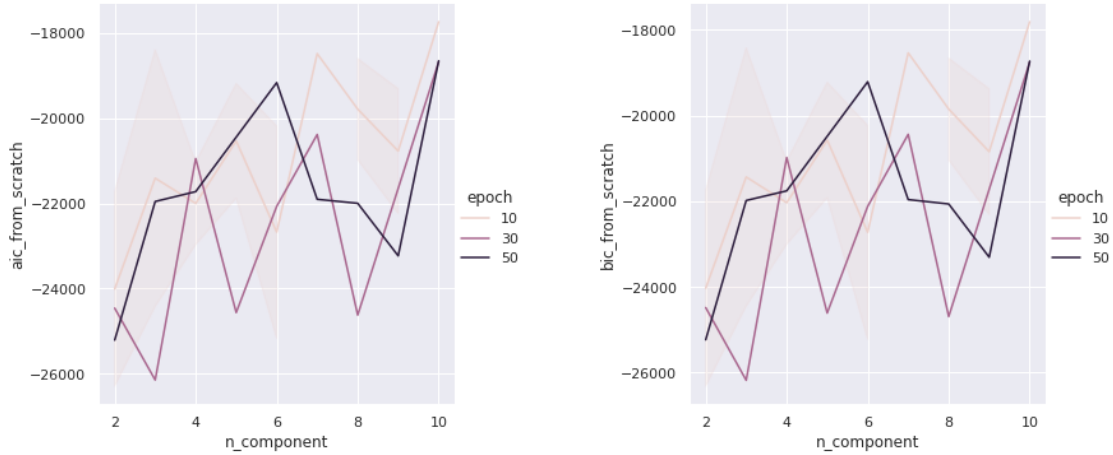
4.3.4 Wyniki dla zbioru opisującego spożycie nikotyny

Podobnie jak w poprzednim rozdziale, pierwszym krokiem było porównanie miar AIC oraz BIC w zależności od różnych wartości hiperparametrów. Wyniki przedstawione zostały na Rysunku 26 dla modelu z biblioteki *sklearn*, oraz na Rysunku 27 dla własnej implementacji modelu. Podsumowując, najlepsze wyniki dla modeli przedstawia Tablica 3. Wyniki są podobne do tych na zbiorze zawierającym spożycie marihuany.



(a) Wartości miary AIC (b) Wartości miary BIC

Rysunek 26: Miary AIC oraz BIC w zależności od liczby iteracji dla modelu z biblioteki *sklearn*



(a) Wartości miary AIC (b) Wartości miary BIC

Rysunek 27: Miary AIC oraz BIC w zależności od liczby iteracji dla własnej implementacji modelu

Model	# iteracji	# komponentów	AIC	BIC
sklearn	10	2	25404.30	26342.72
sklearn	30	2	25404.30	26342.72
sklearn	40	2	25404.30	26342.72
scratch	10	10	-17733.65	-17811.42
scratch	10	3	-18394.83	-18418.15
scratch	10	7	-18482.78	-18537.22

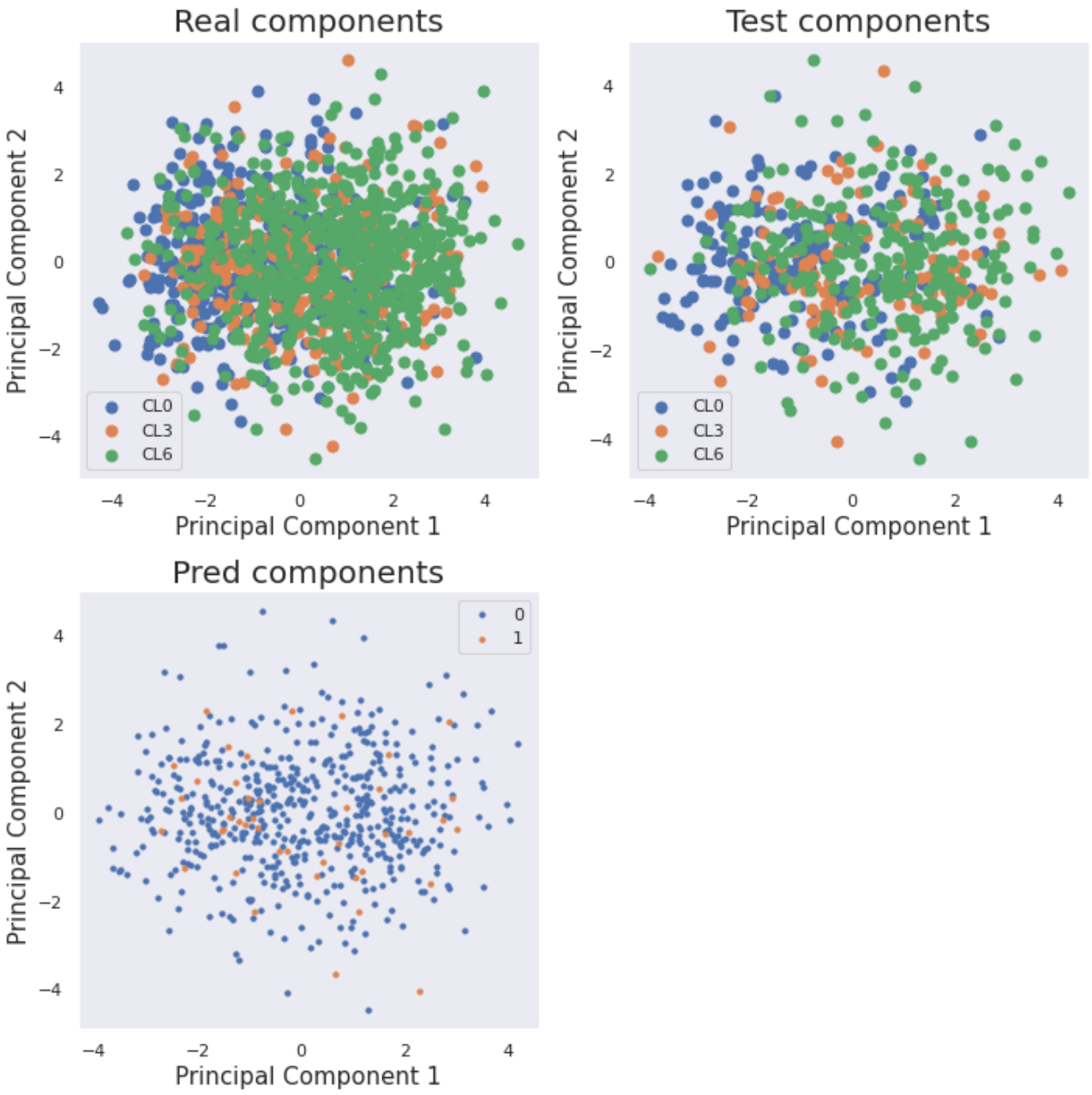
Tablica 3: Najlepsze wartości hiperparametrów względem miar AIC i BIC dla obu modeli

W przypadku modelu z *sklearn* wartości obu metryk są dodatnie, a w przypadku modelu własnego są ujemne. Rozkłady obu metryk pomiędzy modelami się różnią. Dla modelu z *sklearn* widać, że wartości obu metryk maleją wraz ze wzrostem liczby komponentów. Takiej zależności nie widać w przypadku modelu własnej implementacji. Najlepszy wynik osiągnięto dla 10 komponentów, jednak drugi najlepszy z kolei jest dla 3 komponentów (równe liczbie klas w zbiorze). W przypadku modelu z biblioteki liczba iteracji nie ma większego wpływu na wyniki, jednak dla drugiego modelu czym mniej, tym lepiej.

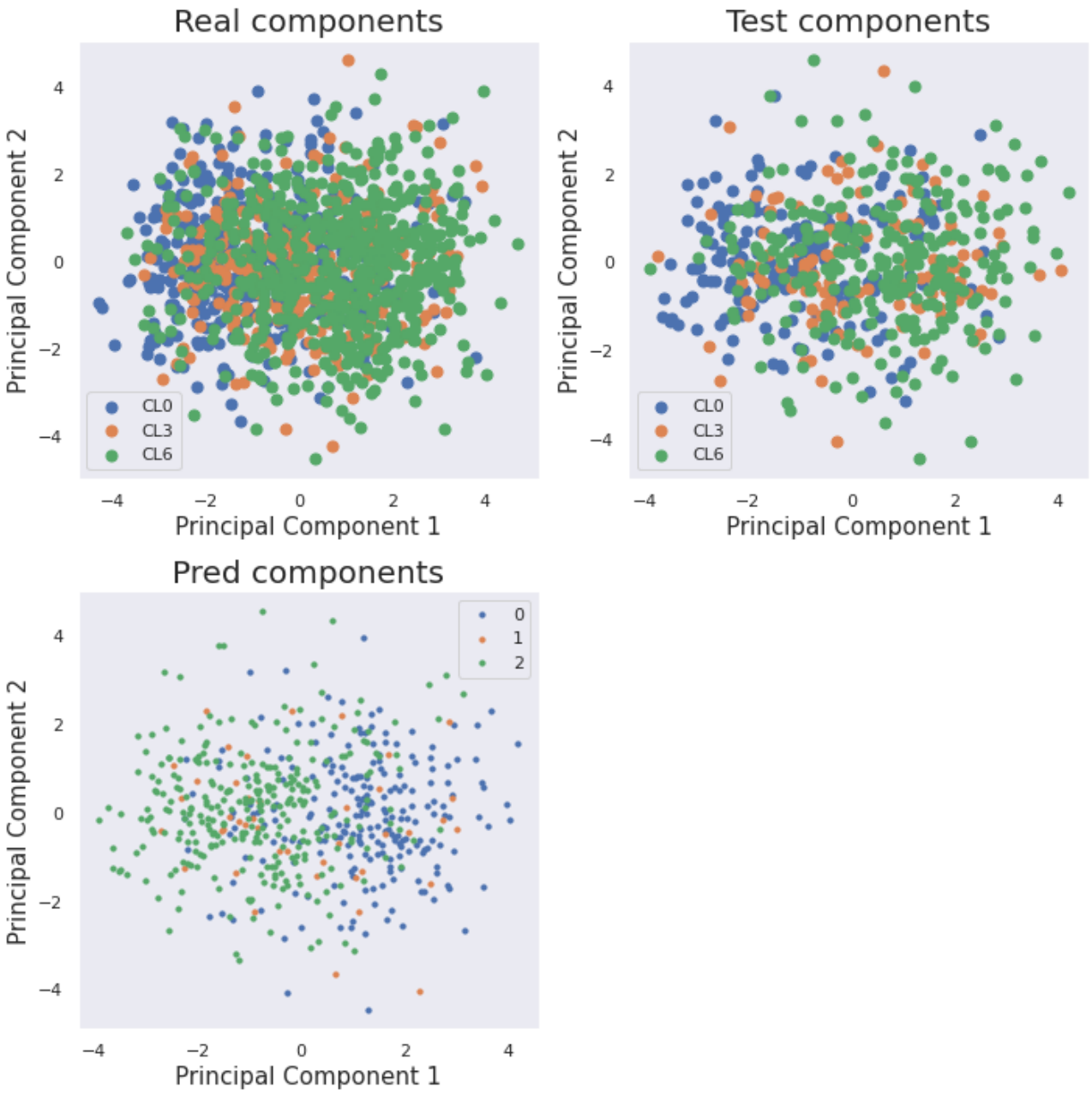
Na Rysunku 34 jak w poprzednim podrozdziale zostały zawarte trzy rysunki. Jednak po bliższym przyjrzeniu się trzeciemu rysunkowi, który odpowiada za predykcję modelu na zbiorze testowym, ciężko jest znaleźć jakiś podział. Większość elementów została zaklasyfikowana do klastra 0. Może to być spowodowane tym, że w zbiorze istnieje więcej niż 2 różne etykiety. Dlatego warto przyjrzeć się podziałowi na 3 klastry, który przedstawiony jest na Rysunku 29. Podział jest dużo bardziej widoczny. Można wyszczególnić trzy klastry, które kolejno odpowiadają etykietom CL0, CL3, CL6. Wniosek jest taki, że miary AIC oraz BIC nie zawsze poprawnie określają najlepszy podział danych i zawsze warto sobie to zwizualizować.

Na Rysunku 30 przedstawiono wizualizację najlepszych hiperparametrów dla modelu implementacji własnej. Podobnie jak ostatnio, najlepsze wyniki osiągnięto dla 10 klastrów i 10 iteracji. Kolor niebieski, czyli klaster o numerze 9, na wykresie predykcji pojawia się prawie w każdym miejscu wykresu. Nie można go więc zidentyfikować z żadną klasą. Inaczej jest w przypadku reszty. Klastry 4 i 8 odpowiadają etykiecie CL0, klastry 0, 5, 6, 7 etykiecie CL3, a klastry 1 oraz 3 etykiecie CL6.

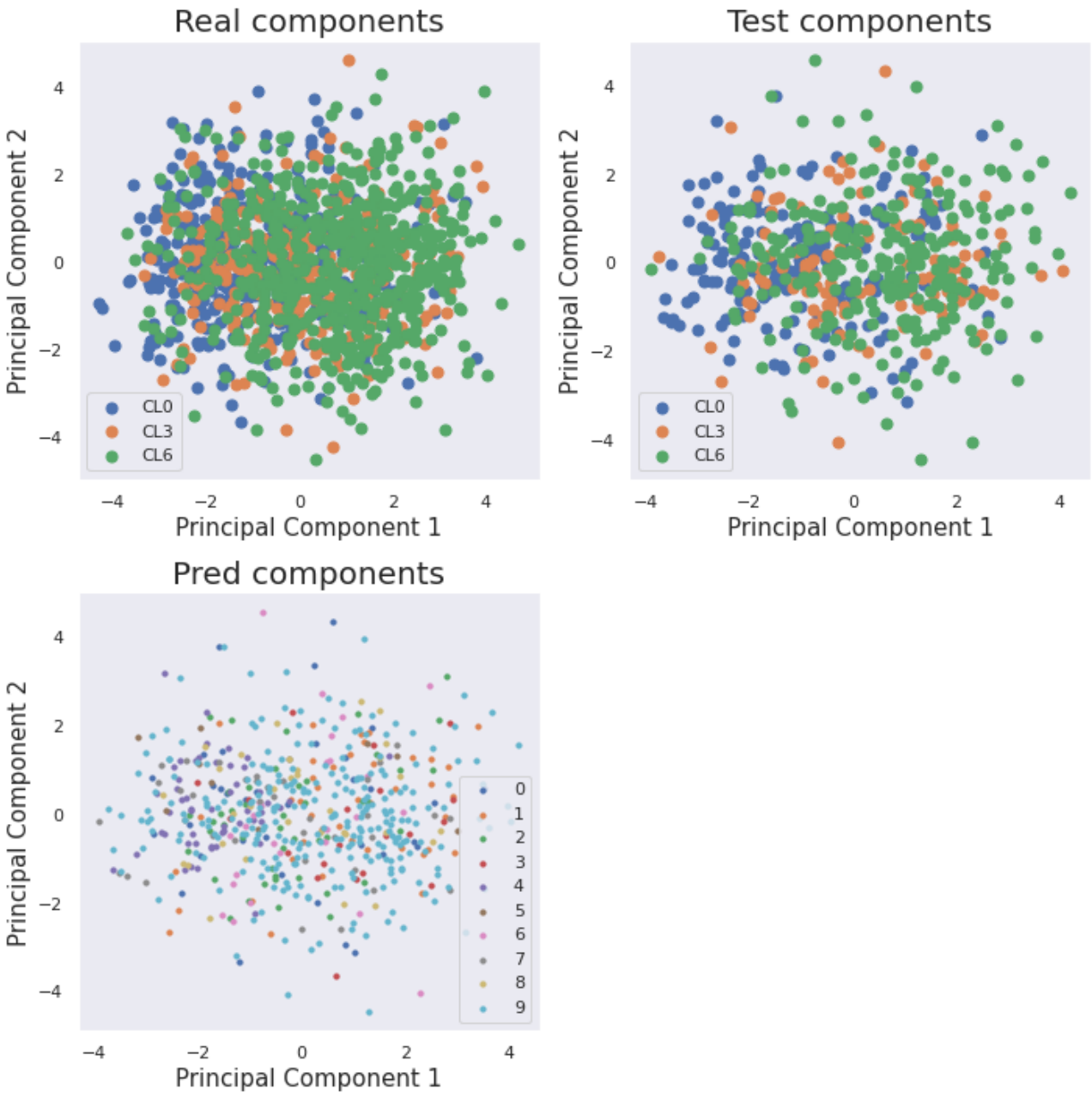
Podobnie jak w poprzednim przypadku, model implementacji własnej lepiej radzi sobie z podziałem na mniejsze klastry, których połączenie tworzy klasy. Model z biblioteki wbudowanej *sklearn* w tym przypadku osiągnął największy likelihood dla podziału, który nie odzwierciedla podziału etykiet. Jednak zmieniając hiperparametry, model całkiem dobrze odzwierciedlił prawdziwe klasy w zbiorze testowym.



Rysunek 28: Wizualizacja danych dla najlepszych hiperparametrów modelu z biblioteki *sklearn*



Rysunek 29: Wizualizacja danych dla 3 klas i 10 iteracji modelu z biblioteki *sklearn*



Rysunek 30: Wizualizacja danych dla najlepszych hiperparametrów modelu własnej implementacji

4.4 Porównanie wyników wybranych modeli

Naiwny Bayes

Najlepsze wyniki uzyskane dla poszczególnych zbiorów w zależności od wykorzystanych cech dla naiwnego Bayesa:

	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*

Rysunek 31: Metryki dla zbioru opisującego spożycie marihuany w zależności od wykorzystanych cech

	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*

Rysunek 32: Metryki dla zbioru opisującego spożycie nikotyny w zależności od wykorzystanych cech

Wyniki są dosyć satysfakcjonujące - szczególnie w przypadku pierwszego zbioru, gdzie udało się uzyskać wyniki F1 macro na poziomie 0.67.

GMM

W przypadku modeli GMM, w celu porównania się z drugą metodą, własnoręcznie dokonano klasyfikacji. Polegała ona na wizualizacji danych oraz na identyfikacji klastrów z klasami. Dzięki temu możliwe było

obliczenie metryk, które przedstawione zostały na Rysunku 33 oraz Rysunku 34.

	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

(a) Metryki modelu z biblioteki *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

(b) Metryki modelu własnej implementacji

Rysunek 33: Metryki dla zbioru opisującego spożycie marihuanny w zależności od modelu

	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

(a) Metryki modelu z biblioteki *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

(b) Metryki modelu własnej implementacji

Rysunek 34: Metryki dla zbioru opisującego spożycie nikotyny w zależności od modelu

Widać, że wyniki nie są dobre. Trochę lepiej z klasyfikacją radzi sobie model z *sklearna*, jednak w każdym przypadku F1 score jest mniejszy niż 0.5.

Wnioski

W przypadku zadania klasyfikacji rozwiązanie oparte o naiwnego Bayesa poradziło sobie zdecydowanie lepiej niż podejście wykorzystujące GMM. Można zauważyć, że oba podejścia miały pewne problemy z ropoznawaniem klasy najsłabiej reprezentowanej, ale poza tym naiwny Bayes dosyć dobrze poradził sobie z pozostałymi klasami, podczas gdy GMM również z nimi miało zauważalne problemy. Jednoznacznie można więc uznać, iż w przypadku badanego problemu naiwny Bayes jest lepszym rozwiązaniem.

5 Podsumowanie

Wykonanie projektu umożliwiło zapoznanie się z całym procesem związanym z wnioskowaniem na podstawie danych - od analizy eksploracyjnej danych, przez ich przetwarzanie i tworzenie modeli uczenia probabilistycznego.

Uzyskane rezultaty pokazały, że w zadaniu klasyfikacji stosowania nikotyny i marihuany lepiej sprawdził się model Naiwnego Bayesa. Może to być spowodowane zarówno przez dane, jak i przez charakterystykę modeli. Modele mikstur rozkładów normalnych głównie służą do klasteryzacji bądź generowania danych, które pochodzą z rozkładów normalnych. W rozpatrywanych danych niestety nie było tak widocznych zależności jak chociażby w przypadku zbiorów analizowanych w trakcie zajęć laboratoryjnych. Jednak mimo wszystko można było zinterpretować jego działanie i zobaczyć jak działa na danych niepochodzących z prawidłowego rozkładu.