Sprawozdanie 4

Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner, 277508

2025-06-22

Spis treści

1 Zadanie 1 **2**
1.1 a) **Wybór i zapoznanie się z danymi** 2

2 zadanie 2 **2**
2.1 a) Wybór i przygotowanie danych 2
2.2 b) Grupowanie i wizualizacja 4
2.3 c) Ocena jakości grupowania. Wybór optymalnej liczby skupień i porównanie metod. 8

1 Zadanie 1

1.1 a) Wybór i zapoznanie się z danymi

Opis zmiennych w zbiorze danych **Wine**

Kolumna	Nazwa zmiennej	Opis
V1	Alcohol	Zawartość alkoholu (%)
V2	Malic acid	Zawartość kwasu jabłkowego (g/l)
V3	Ash	Zawartość popiołu (g/l)
V4	Alcalinity of ash	Zasadowość popiołu (g/l)
V5	Magnesium	Zawartość magnezu (mg/l)
V6	Total phenols	Zawartość fenoli ogółem (g/l)
V7	Flavanoids	Zawartość flawonoidów (g/l)
V8	Nonflavanoid phenols	Zawartość fenoli nienależących do flawonoidów (g/l)
V9	Proanthocyanins	Zawartość proantocyjaninów (g/l)
V10	Color intensity	Intensywność koloru (od 0 do 13)
V11	Hue	Odcień barwy
V12	OD280/OD315 of diluted wines	Absorbancja przy długości fali 280 nm do 315 nm (rozcieńczone wino)
V13	Proline	Zawartość proliny (mg/l)

2 zadanie 2

2.1 a) Wybór i przygotowanie danych

Do porównywania metod grupujących i hierarchicznych wybieramy dane Sonar

Tutaj przykładowe pierwsze 5 wierszy i 5 kolumn z danych Sonar

```
#DANE
#-----------------------------
data("Sonar")
dane <- Sonar
head(dane[1:5,1:5])
```

```
##       V1     V2     V3     V4     V5
## 1 0.0200 0.0371 0.0428 0.0207 0.0954
## 2 0.0453 0.0523 0.0843 0.0689 0.1183
## 3 0.0262 0.0582 0.1099 0.1083 0.0974
## 4 0.0100 0.0171 0.0623 0.0205 0.0205
## 5 0.0762 0.0666 0.0481 0.0394 0.0590
```

```
#-----------------------------
```

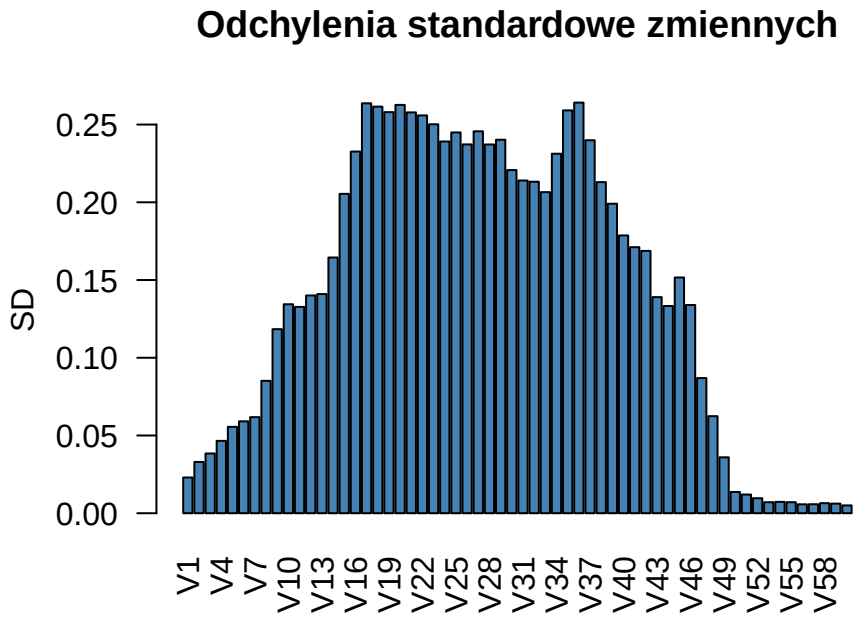
Dane mają 208 wierszy oraz 61 kolumn, więc nie jest potrzebne losowanie 200 wierszy z 208,

nie zmniejszy to w znaczący sposób zapotrzebowania pamięciowego programu.

Następnie usuwamy zmienną grupującą

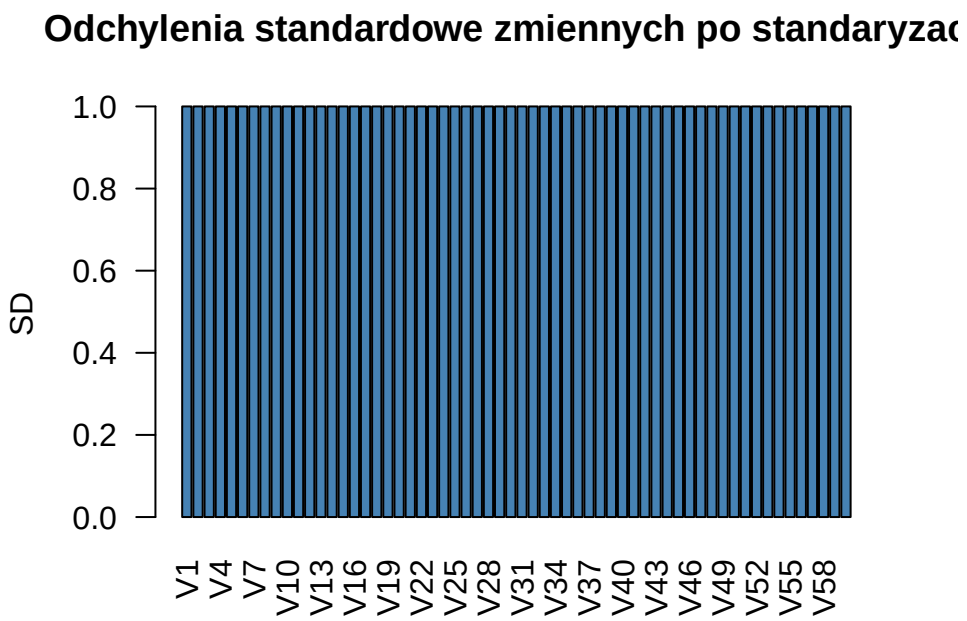
```
#USUWANIE KOLUMNY Z ETYKIETAMI
#-----------------------------
Y <- dane[,61]
X <- dane[,-61]
#-----------------------------
```

Teraz trzeba rozstrzygnąć czy konieczne jest zastosowanie standaryzacji przed wyznaczeniem macierzy odległości/odmienności



Widać duże różnice między odchyleniami standardowymi, co zmusza nas do wprowadzenia standaryzacji

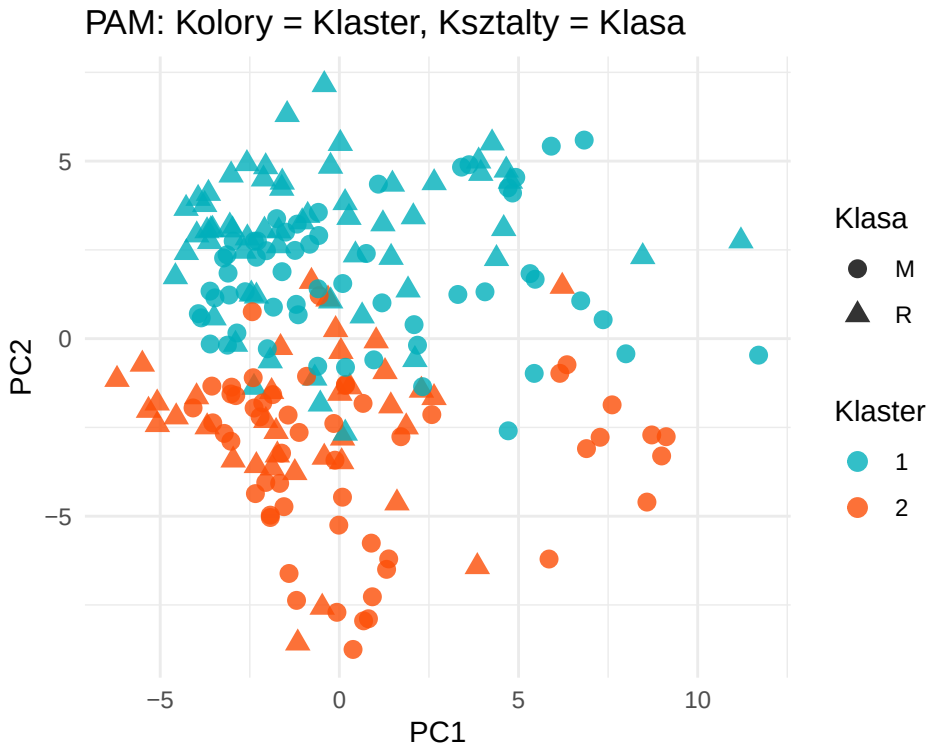
Przeprowadzamy standaryzację, i po niej odchylenia standardowe są równe



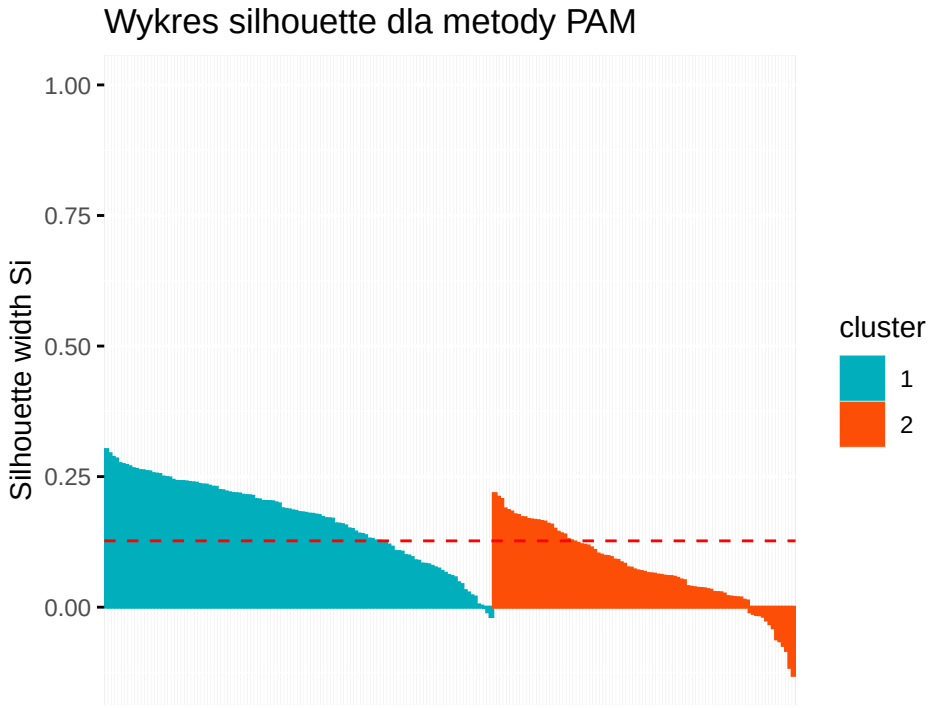
Po standaryzacji odchylenia standardowe wszystkich zmiennych są sobie równe

2.2 b) Grupowanie i wizualizacja

Przeprowadzamy grupowanie za pomocą metody PAM




```
##   cluster size ave.sil.width
## 1       1  117          0.17
## 2       2   91          0.07
```



Z analizy wykresu wynika, że uzyskane skupienia charakteryzują się umiarkowaną separacją, co wskazuje na częściowe nakładanie się klastrów. Pomimo tego, klastry wykazują dobrą zwartość, co oznacza, że obiekty w obrębie poszczególnych skupień są do siebie stosunkowo podobne. Dodatkowo, skupienia cechują się względną jednorodnością, co sugeruje spójną strukturę wewnątrz każdego klastra.

Tabela 2: Tabela zgodności: klastry PAM vs klasy rzeczywiste

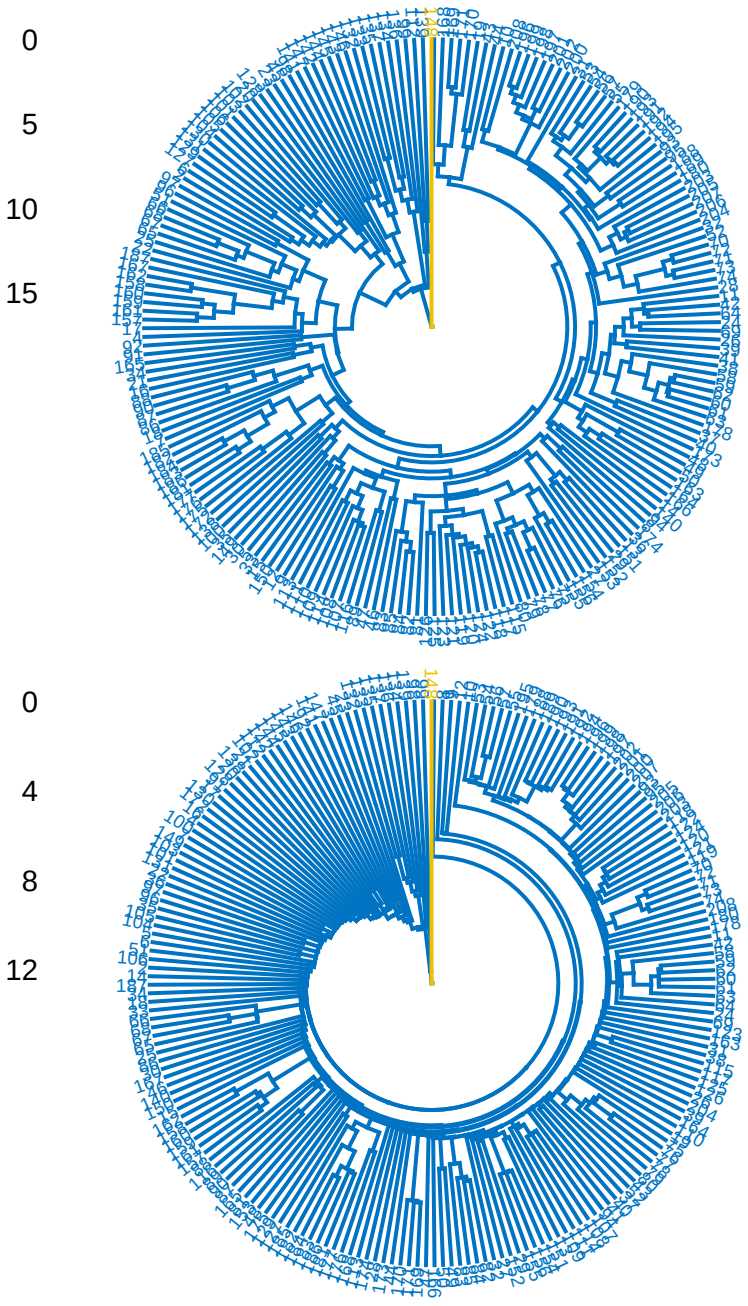
M	R
57	60
54	37

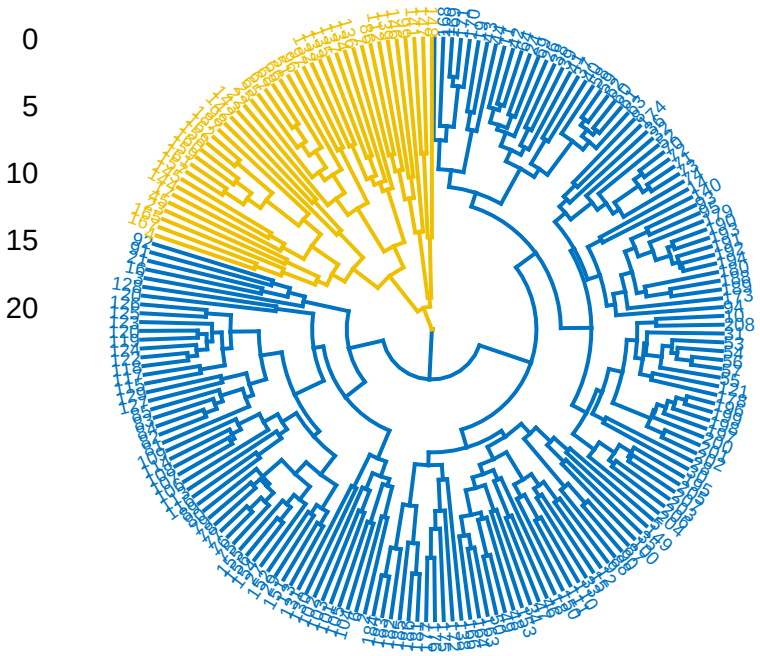
```
## Dokładność przypisania (accuracy): 54.81 %
```

```
## Adjusted Rand Index (ARI): 0.004
```

Niestety zgodnośc w wyniku takiego przypisania wynosi jedynie 55%, jest to zgodność na bardzo słabym poziomie, więc nie jest dobrym pomysłem, używanei tego typu sposobu grupwoanai do przedstawionych danych

Podział hierarchiczny


0
5
10
15
0
4
8
12



Na dendrogramach zaobserwować można występowanie efektu tzw. „łańcucha” zarówno w metodach single linkage, jak i average linkage. Metoda single linkage łączy klastry na podstawie minimalnej odległości pomiędzy pojedynczymi punktami z różnych grup. W sytuacji, gdy w zbiorze danych znajduje się punkt znacznie odległy od pozostałych, może on powodować powstawanie rozciągniętych, „łańcuchowatych” struktur, co objawia się efektem łańcucha.

Analiza wykresu rozrzutu (scatterplot) w przypadku grupowania metodą PAM ukazuje wyraźnie oddalone punkty po prawej stronie, które właśnie stanowią przyczynę obserwowanego efektu łańcucha w dendrogramach.

Analogiczna sytuacja zachodzi w metodzie average linkage, gdzie odległość między klastrami definiowana jest jako średnia odległość wszystkich par punktów pomiędzy dwoma klastrami. Obecność pojedynczego, silnie oddalonego punktu powoduje, że średnia odległość tego punktu do pozostałych klastrów jest wysoka, co skutkuje jego izolacją jako osobnego klastra lub bardzo spolaryzowanym przyporządkowaniem w hierarchii.

W praktyce oznacza to, że zarówno single linkage, jak i average linkage mogą być podatne na wpływ punktów odstających (outliers), co należy uwzględnić podczas interpretacji wyników klasteryzacji.

Metoda complete linkage (łączenie przez najdalszego sąsiada) definiuje odległość między dwoma klastrami jako maksymalną odległość pomiędzy dowolnymi parami punktów z tych klastrów. Oznacza to, że dwa klastry zostaną połączone tylko wtedy, gdy każdy punkt jednego klastra znajduje się stosunkowo blisko każdego punktu drugiego klastra.

W praktyce przekłada się to na tworzenie bardziej zwartych i zrównoważonych grup, co widać również w naszym przypadku — około 25% obserwacji zostało przypisanych do jednej grupy, a pozostałe 75% do drugiej, co świadczy o lepszym rozdzieleniu klastrów i mniejszym wpływie pojedynczych, odległych punktów na strukturę grup.

2.3 c) Ocena jakości grupowania. Wybór optymalnej liczby skupień i porównanie metod.

Tabela 3: Procentowa zgodność grupowania dla różnych metod i liczby klastrów

	Liczba_klastrow	PAM	AGNES_avg	AGNES_single	AGNES_complete
2	2	54.81	53.37	53.37	53.37
3	3	58.65	53.37	53.37	53.37
4	4	57.69	53.37	53.37	53.37
5	5	67.31	54.33	53.37	56.73
6	6	67.31	54.81	53.37	58.17
7	7	67.79	54.81	53.85	60.58
8	8	67.79	58.17	54.33	63.46
9	9	67.31	58.65	54.81	63.46
10	10	67.79	59.13	54.81	63.46