

# Raport - Zaawansowane metody klasyfikacji oraz analiza skupień – algorytmy grupujące i hierarchiczne

Filip Michewicz 282239  
Wiktor Niedźwiedzki 258882

10 czerwca 2025 Anno Domini

## Spis treści

<b>1</b>	<b>Zaawansowane metody klasyfikacji</b>	<b>3</b>
1.1	Rodziny klasyfikatorów/uczenie zespołowe . . . . .	3
1.1.1	Drzewa klasyfikacyjne . . . . .	3
1.2	Metoda wektorów nośnych (SVM) . . . . .	3
1.2.1	Jądro liniowe . . . . .	3
1.2.2	Jądro wielomianowe . . . . .	4
1.2.3	Jądro radialne . . . . .	6
1.2.4	Jądro sigmoidalne . . . . .	7
1.3	Wnioski . . . . .	9
<b>2</b>	<b>Analiza skupień – algorytmy grupujące i hierarchiczne</b>	<b>9</b>
2.1	Zbiór danych Glass . . . . .	9
2.2	Wyniki grupowania . . . . .	10
2.2.1	k-średnie . . . . .	11
2.2.2	Partitioning Around Medoids (PAM) . . . . .	12
2.2.3	Agglomerative Nesting (AGNES) . . . . .	13
2.2.4	Divisive clustering (DIANA) . . . . .	16
2.3	Ocena jakości grupowania i wizualizacja najlepszych wyników . . . . .	18
2.3.1	Ocena . . . . .	18
2.3.2	Wizualizacja . . . . .	20
2.4	Wnioski . . . . .	25
<b>3</b>	<b>Podsumowanie</b>	<b>25</b>

## Spis wykresów

1	Wykres pudełkowy, zmienne bez standaryzacji . . . . .	9
2	Wykres pudełkowe, po standaryzacji . . . . .	10
3	Wizualizacja danych, PCA . . . . .	10
4	PCA, kolory - rzeczywiste, kształt - wyniki . . . . .	11
5	Wykres RI od Na, aby pokazać gdzie są wyznaczone centra skupień . . . . .	11
6	coś . . . . .	12
7	coś, z medoidami . . . . .	13
8	AGNES: single linkage . . . . .	14
9	AGNES: complete linkage . . . . .	15
10	AGNES: average linkage . . . . .	16
11	AGNES: average linkage . . . . .	17

12	AGNES: average linkage . . . . .	18
13	Connectivity . . . . .	19
14	Dunn . . . . .	19
15	Silhouette . . . . .	20
16	PCA, kolory - rzeczywiste, kształt - wyniki, k=2 . . . . .	21
17	Wykres RI od Na, aby pokazać gdzie są wyznaczone centra skupień, k=2 . . . . .	21
18	coś, k=2 . . . . .	22
19	coś, z medoidami, k=2 . . . . .	22
20	AGNES: complete lineage, k=2 . . . . .	24
21	DIANA, k=3 . . . . .	25

## Spis tabel

1	Średnia poprawa dokładności klasyfikacji za pomocą drzewa klasyfikacyjnego, z podziałem na algorytmny uczenia zespołowego oraz liczbę replikacji . . . . .	3
2	Jądro liniowe - bez skalowania . . . . .	3
3	Jądro liniowe - ze skalowaniem . . . . .	4
4	Jądro wielomianowe - wielokrotny podział, bez skalowania . . . . .	4
5	Jądro wielomianowe - wielokrotny podział, ze skalowaniem . . . . .	4
6	Jądro wielomianowe - cross-validation, bez skalowania . . . . .	4
7	Jądro wielomianowe - cross-validation, ze skalowaniem . . . . .	5
8	Jądro wielomianowe - bootstrap, bez skalowania . . . . .	5
9	Jądro wielomianowe - bootstrap, ze skalowaniem . . . . .	5
10	Badanie wpływu stopnia wielomianu na dokładność - wielokrotny podział, najbardziej dokładna kombinacja gammy i kary dla opcji default (stopień 3 . . . . .	6
11	Jądro radialne - wielokrotny podział, bez skalowania . . . . .	6
12	Jądro radialne - wielokrotny podział, ze skalowaniem . . . . .	6
13	Jądro radialne - cross-validation, bez skalowania . . . . .	6
14	Jądro radialne - cross-validation, ze skalowaniem . . . . .	7
15	Jądro radialne - bootstrap, bez skalowania . . . . .	7
16	Jądro radialne - bootstrap, ze skalowaniem . . . . .	7
17	Jądro sigmoidalne - wielokrotny podział, bez skalowania . . . . .	7
18	Jądro sigmoidalne - wielokrotny podział, ze skalowaniem . . . . .	8
19	Jądro sigmoidalne - cross-validation, bez skalowania . . . . .	8
20	Jądro sigmoidalne - cross-validation, ze skalowaniem . . . . .	8
21	Jądro sigmoidalne - bootstrap, bez skalowania . . . . .	8
22	Jądro sigmoidalne - bootstrap, ze skalowaniem . . . . .	9
23	Macierz błędów; metoda k-średnich . . . . .	11
24	Dane medoidów, k=6 . . . . .	13
25	Macierz błędów; metoda k-średnich . . . . .	13
26	Macierz błędów; agnes, najbliższy sąsiad . . . . .	14
27	Macierz błędów; agnes, najdalszy sąsiad . . . . .	15
28	Macierz błędów; agnes, średnia odległość . . . . .	16
29	Macierz błędów; agnes, średnia odległość . . . . .	17
30	Macierz błędów; agnes, średnia odległość . . . . .	18
31	Dane medoidów, k=2 . . . . .	22

# 1 Zaawansowane metody klasyfikacji

W pierwszej części zadania zastosujemy algorytmy *ensemble learning* (bagging, boosting i random forest) w celu poprawy dokładności cech klasyfikacyjnych. W drugiej natomiast poznamy i ocenimy nową metodę klasyfikacji - metodę wektorów nośnych (SVM).

Zadanie zostanie wykonane na zbiorze danych *wine*, którego szczegółowy opis znajduje się w poprzednim raporcie.

## 1.1 Rodziny klasyfikatorów/uczenie zespołowe

Wyróżniamy trzy algorytmy uczenia zespołowego (ang. ensemble learning):

- **Bagging** - generujemy B-bootstrapowych replikacji zbioru uczącego, na podstawie których tworzymy B klasyfikatorów. Następnie łączymy je w klasyfikator zagregowany, który przydziela dane cechy do klas za pomocą reguły "głosowania większości" (w przypadku remisu wybiera losowo). Każdy klasyfikator powstaje niezależnie (w sensie takim, że wyniki poprzednich nie mają wpływu na generowanie nowych).
- **boosting** - podobnie jak w bagging, tworzymy klasyfikator zagregowany złożony z wielu pojedynczych klasyfikatorów. Jednak różnica jest taka, że klasyfikatory powstają sekwencyjnie. Na początku każda cecha w zbiorze ma przypisaną taką samą wagę. Z każdą kolejną iteracją natomiast waga zwiększa się dla uprzednio źle sklasyfikowanych przypadków.
- **random forest** (dla drzew klasyfikacyjnych) - metoda podobna do bagging z tą różnicą, że klasyfikatory powstają na podstawie różnych m-elementowych podzbiorach cech (m mniejsze bądź równe wszystkim cechom).

### 1.1.1 Drzewa klasyfikacyjne

#### MOŻE BYĆ NIEPOOPRAWNIE W CHUJ

Tabela 1: Średnia poprawa dokładności klasyfikacji za pomocą drzewa klasyfikacyjnego, z podziałem na algorytmy uczenia zespołowego oraz liczbę replikacji

	1	5	10	20	30	40	50	100
Bagging	19.08	52.50	47.25	64.91	61.25	51.00	66.50	57.71
Random Forest	87.44	88.16	87.55	84.90	84.69	86.26	87.59	84.90
Boosting	73.49	70.44	73.56	74.10	79.94	75.98	75.62	67.32
Średnia	60.01	70.37	69.45	74.64	75.29	71.08	76.57	69.98

## 1.2 Metoda wektorów nośnych (SVM)

W tej części przeprowadzona będzie klasyfikacja na podstawie metody wektorów nośnych, z podziałem na różne funkcje jądrowe.

#### COŚ O TYM CO TO WOGÓLE JEST

### 1.2.1 Jądro liniowe

Tabela 2: Jądro liniowe - bez skalowania

	0.001	0.01	0.1	1	10	100	1000
Wielokrotny podział	37.00	97.17	97.67	96.00	97.33	96.50	97.50
Cross-validation	33.33	50.12	90.07	89.51	89.51	89.51	89.51
Bootstrap	38.32	96.95	97.17	95.68	96.50	97.24	96.49

Tabela 2: Jądro liniowe - bez skalowania (kontynuacja)

	0.001	0.01	0.1	1	10	100	1000
Średnio	36.22	81.41	94.97	93.73	94.45	94.42	94.50

Tabela 3: Jądro liniowe - ze skalowaniem

	0.001	0.01	0.1	1	10	100	1000
Wielokrotny podział	38.33	96.50	96.17	95.50	96.17	96.67	96.17
Cross-validation	44.39	49.70	92.19	92.51	92.51	92.51	92.51
Bootstrap	40.23	96.74	96.52	96.76	96.66	96.67	96.78
Średnio	40.99	80.98	94.96	94.92	95.11	95.28	95.15

### 1.2.2 Jądro wielomianowe

Tabela 4: Jądro wielomianowe - wielokrotny podział, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	35.67	94.83	95.67	92.50	96.17	95.50	95.17	96.67	95.17	94.83
0.01	35.50	93.83	94.00	96.00	95.50	94.50	96.17	94.17	94.33	95.17
0.1	41.00	95.50	96.83	94.67	95.33	95.33	95.50	93.50	94.33	95.50
1	35.83	95.17	94.83	92.83	95.00	96.00	94.83	95.67	95.33	94.17
10	41.67	95.83	95.33	96.50	95.17	94.67	95.67	95.17	96.33	94.50
100	86.00	95.33	96.67	94.00	94.83	94.67	95.00	95.17	95.33	95.33
1000	95.67	95.33	95.67	96.17	96.50	96.17	96.17	97.00	95.33	95.50

Tabela 5: Jądro wielomianowe - wielokrotny podział, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	38.17	96.00	94.67	96.33	95.33	96.00	95.33	95.50	94.17	93.83
0.01	39.67	95.83	95.17	96.50	94.00	96.33	95.67	94.83	96.33	96.83
0.1	35.83	95.83	96.50	96.33	94.67	95.17	94.83	95.33	95.17	95.33
1	38.17	94.00	96.83	94.83	95.67	94.67	96.67	94.67	96.00	95.33
10	40.00	93.83	96.17	95.33	96.00	95.00	95.67	96.00	96.50	95.17
100	87.67	96.17	95.33	95.83	95.83	96.17	95.50	96.00	96.17	94.50
1000	96.33	94.67	96.00	94.00	94.83	95.50	95.00	95.00	95.00	94.83

Tabela 6: Jądro wielomianowe - cross-validation, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	39.80	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11
0.01	39.80	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11
0.1	39.80	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11
1	39.80	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11
10	44.28	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11

Tabela 6: Jądro wielomianowe - cross-validation, bez skalowania  
(kontynuacja)

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
100	87.61	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11
1000	95.00	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11	96.11

Tabela 7: Jądro wielomianowe - cross-validation, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	40.00	97.78	96.08	96.67	96.67	96.67	96.67	96.67	96.67	96.67
0.01	40.00	96.08	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
0.1	40.00	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
1	40.00	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
10	43.95	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
100	88.24	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
1000	96.11	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67

Tabela 8: Jądro wielomianowe - bootstrap, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	37.30	94.96	92.79	93.79	93.06	96.20	95.14	94.33	92.90	93.78
0.01	36.30	94.38	93.73	94.72	95.31	94.64	94.58	94.85	94.79	93.92
0.1	36.85	93.90	94.55	93.91	95.05	93.28	94.25	94.46	94.68	93.72
1	32.77	93.92	95.03	94.77	92.97	94.20	93.11	93.36	94.73	94.13
10	39.36	95.14	95.54	96.42	93.46	94.80	93.71	94.34	94.90	92.69
100	86.17	95.19	94.79	94.86	92.67	95.01	94.67	93.97	93.45	95.10
1000	94.38	94.81	93.22	93.29	94.01	94.91	93.04	94.51	94.13	94.48

Tabela 9: Jądro wielomianowe - bootstrap, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	35.97	93.62	94.60	94.90	93.00	93.75	94.94	92.08	94.30	94.27
0.01	37.09	94.76	94.26	94.54	94.29	95.00	95.70	93.39	95.81	93.90
0.1	39.87	95.09	94.28	93.77	94.31	92.36	94.92	94.60	95.24	94.70
1	32.07	93.55	94.86	94.43	93.36	93.91	94.14	94.54	95.61	93.10
10	39.96	94.39	94.78	94.43	94.73	94.75	93.79	94.75	95.14	95.35
100	87.85	95.95	96.22	94.10	93.82	94.00	93.64	94.64	93.62	94.45
1000	93.80	94.77	96.07	94.57	95.15	94.99	92.96	95.04	94.78	92.77

Najlepsza gamma: **10**, najlepsza kara: **0.01**. Robimy dla danych po standaryzacji, bo tak i chuj.  
Badamy tylko na podstawie wielokrotnego podziału, bo tak i chuj również.

Tabela 10: Badanie wpływu stopnia wielomianu na dokładność - wielokrotny podział, najbardziej dokładna kombinacja gammy i kary dla opcji default (stopień 3

	2	3	4	5	6	7
Dokładność	88.33	95.17	88.67	89	77.83	79.17

### 1.2.3 Jądro radialne

Tabela 11: Jądro radialne - wielokrotny podział, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03
0.01	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03
0.1	80.85	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03
1	98.30	56.83	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03
10	96.63	59.61	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03
100	96.63	59.61	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03
1000	96.63	59.61	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03

Tabela 12: Jądro radialne - wielokrotny podział, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
0.01	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
0.1	80.85	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
1	98.30	53.86	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
10	97.75	59.51	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
100	97.19	59.51	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
1000	97.19	59.51	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87

Tabela 13: Jądro radialne - cross-validation, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	39.80	39.80	39.8	39.8	39.8	39.8	39.8	39.8	39.8	39.8
0.01	39.80	39.80	39.8	39.8	39.8	39.8	39.8	39.8	39.8	39.8
0.1	79.74	39.80	39.8	39.8	39.8	39.8	39.8	39.8	39.8	39.8
1	98.89	57.84	39.8	39.8	39.8	39.8	39.8	39.8	39.8	39.8
10	97.22	61.76	39.8	39.8	39.8	39.8	39.8	39.8	39.8	39.8
100	95.56	61.76	39.8	39.8	39.8	39.8	39.8	39.8	39.8	39.8
1000	95.56	61.76	39.8	39.8	39.8	39.8	39.8	39.8	39.8	39.8

Tabela 14: Jądro radialne - cross-validation, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97
0.01	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97
0.1	80.92	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97
1	98.30	57.45	39.97	39.97	39.97	39.97	39.97	39.97	39.97	39.97
10	97.19	61.44	40.56	39.97	39.97	39.97	39.97	39.97	39.97	39.97
100	96.08	61.44	40.56	39.97	39.97	39.97	39.97	39.97	39.97	39.97
1000	96.08	61.44	40.56	39.97	39.97	39.97	39.97	39.97	39.97	39.97

Tabela 15: Jądro radialne - bootstrap, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	39.44	35.19	38.09	37.56	38.24	36.51	37.37	33.77	37.01	38.05
0.01	37.61	38.08	34.79	37.87	35.97	37.30	39.55	36.03	38.57	33.31
0.1	52.32	36.76	33.05	37.36	35.12	36.75	38.88	32.40	40.04	38.82
1	97.13	49.89	37.44	37.13	37.20	40.12	36.40	37.28	38.82	35.95
10	96.85	53.86	38.25	37.49	38.28	36.42	36.90	37.72	38.51	39.46
100	97.51	48.61	41.25	37.40	39.36	36.69	36.11	38.37	37.57	36.84
1000	96.04	51.69	39.35	37.67	40.31	40.71	35.66	36.45	39.61	36.03

Tabela 16: Jądro radialne - bootstrap, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	36.38	38.00	34.28	35.66	38.14	35.34	38.63	38.56	36.59	36.93
0.01	44.29	34.10	36.08	33.77	37.20	35.12	35.16	35.62	38.13	35.21
0.1	51.91	36.81	36.89	34.64	37.43	37.62	35.26	35.74	39.98	37.51
1	97.13	45.91	41.10	37.88	38.31	36.54	40.25	37.96	37.35	36.27
10	96.99	55.21	36.87	38.56	38.27	38.40	37.21	35.04	39.23	37.96
100	97.00	55.27	38.28	38.21	38.07	38.63	38.60	37.05	39.56	36.93
1000	96.72	55.64	35.79	36.66	36.84	35.85	32.43	36.38	38.67	36.04

#### 1.2.4 Jądro sigmoidalne

Tabela 17: Jądro sigmoidalne - wielokrotny podział, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	38.50	38.83	38.50	40.33	41.33	46.33	38.33	39.17	38.83	43.17
0.01	38.83	63.00	57.83	61.83	67.50	68.33	60.67	59.33	62.83	61.50
0.1	40.33	89.50	89.00	87.50	86.33	90.33	88.50	87.17	86.67	86.00
1	97.00	81.17	81.83	81.33	81.50	78.83	81.17	82.00	79.00	82.83
10	98.00	84.17	83.67	78.50	79.83	80.00	81.67	79.67	80.17	80.33
100	96.50	81.50	82.67	79.83	81.83	80.17	80.00	82.33	79.83	79.17
1000	97.17	79.33	79.83	79.83	80.83	79.33	82.33	81.00	77.83	82.50

Tabela 18: Jądro sigmoidalne - wielokrotny podział, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	40.17	39.33	42.50	36.33	45.67	37.50	36.83	37.83	40.50	37.00
0.01	39.00	64.17	63.67	63.00	65.67	64.50	70.83	55.50	70.00	57.00
0.1	38.50	89.33	86.50	86.33	86.83	88.33	85.33	87.17	88.00	87.83
1	98.50	83.00	79.50	79.50	82.33	80.00	78.17	81.83	80.50	83.83
10	98.00	83.67	78.33	82.83	79.17	78.00	81.67	81.00	84.17	82.67
100	96.17	80.67	79.83	81.50	82.17	78.00	78.50	80.33	79.83	80.33
1000	97.17	84.50	78.50	81.50	80.50	78.00	80.33	80.50	78.33	81.00

Tabela 19: Jądro sigmoidalne - cross-validation, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	39.93	39.93	39.93	39.93	39.93	39.93	39.93	39.93	39.93	39.93
0.01	39.93	79.31	78.76	79.28	79.28	79.28	79.28	79.28	78.73	78.73
0.1	42.75	87.03	88.17	87.06	87.06	87.58	87.61	85.39	85.95	85.95
1	97.78	83.82	82.61	80.88	76.41	76.37	76.47	78.14	78.66	80.88
10	99.44	81.54	79.74	80.85	80.36	77.03	76.47	76.54	80.39	80.39
100	96.05	82.65	78.63	79.67	79.22	75.33	74.77	75.95	78.73	80.39
1000	96.60	83.20	80.29	80.23	76.96	74.18	73.63	75.92	79.25	78.73

Tabela 20: Jądro sigmoidalne - cross-validation, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
0.01	39.87	75.75	74.64	75.20	75.20	75.20	75.20	75.20	75.20	75.20
0.1	42.68	88.20	87.03	87.06	87.61	86.50	86.50	87.06	86.50	85.95
1	98.30	81.41	80.88	78.63	79.25	80.36	78.14	77.06	78.73	76.47
10	98.89	77.55	80.98	74.15	76.37	75.85	77.03	74.22	75.92	75.33
100	97.16	81.41	81.47	74.15	75.23	77.48	78.10	77.03	77.61	77.58
1000	97.75	82.03	81.47	76.37	75.26	76.37	77.52	75.33	76.50	77.03

Tabela 21: Jądro sigmoidalne - bootstrap, bez skalowania

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	38.03	35.42	32.60	33.82	37.92	40.67	36.77	36.83	40.69	38.03
0.01	38.76	56.17	55.34	57.54	67.60	62.66	62.88	51.72	57.66	61.78
0.1	35.08	85.15	87.53	86.61	86.67	87.19	87.28	86.17	86.12	84.70
1	96.43	79.20	83.35	79.20	80.82	81.15	81.00	84.18	78.42	80.30
10	97.78	83.04	80.24	76.51	79.92	80.63	80.50	79.17	82.40	81.75
100	97.11	76.26	79.87	78.22	77.83	79.00	77.89	79.94	78.86	83.10
1000	95.41	76.93	80.35	80.86	79.87	77.75	78.71	77.39	79.76	78.75



Tabela 22: Jądro sigmoidalne - bootstrap, ze skalowaniem

	0.01	1.12	2.23	3.34	4.45	5.56	6.67	7.78	8.89	10
0.001	35.96	33.55	36.45	38.58	40.31	41.96	39.79	33.05	42.64	38.63
0.01	44.24	56.08	58.36	57.75	65.58	64.09	62.02	56.66	70.28	60.61
0.1	37.47	88.41	86.62	85.84	85.19	88.23	84.05	84.71	84.97	84.48
1	96.89	83.02	78.30	77.96	81.46	79.68	78.83	81.34	80.43	79.81
10	96.90	80.66	78.31	79.23	78.28	78.48	83.28	80.58	79.72	79.16
100	95.16	81.85	77.85	77.58	82.50	82.16	82.94	80.50	77.31	79.11
1000	96.86	80.48	80.60	80.61	81.92	80.41	82.63	80.72	82.08	79.93

elo

### 1.3 Wnioski

e

## 2 Analiza skupień – algorytmy grupujące i hierarchiczne

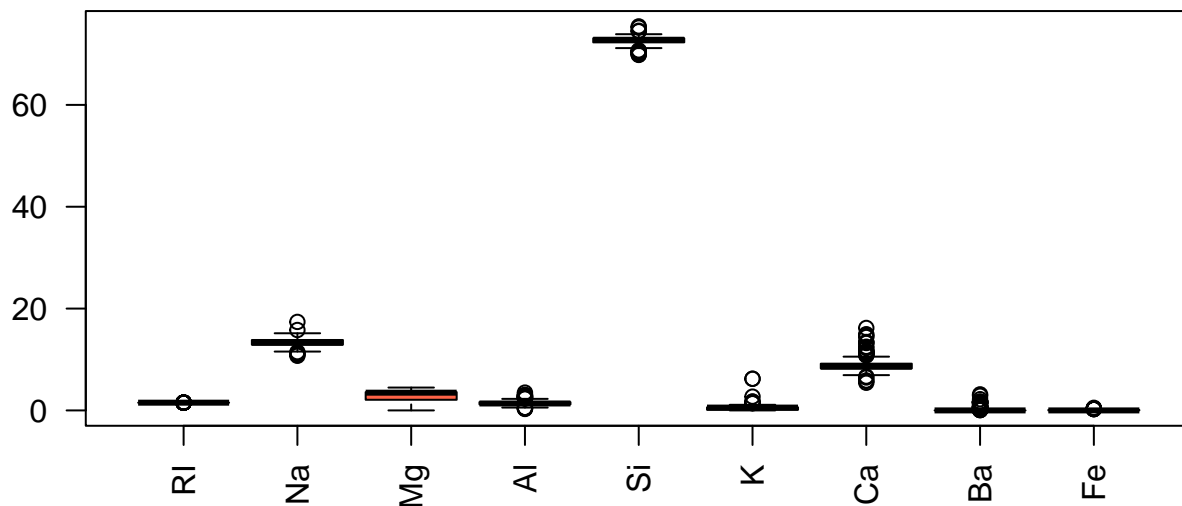
W pierwszej części zadania zastosujemy algorytmy *ensemble learning* (bagging, boosting i random forest) w celu poprawy dokładności cech klasyfikacyjnych. W drugiej natomiast poznamy i ocenimy nową metodę klasyfikacji - metodę wektorów nośnych (SVM).

Zadanie zostanie wykonane na zbiorze danych *wine*, którego szczegółowy opis znajduje się w poprzednim raporcie.

W tym zadaniu zastosujemy i porównamy ze sobą metody analizy skupień - k-średnich i PAM jako algorytmy grupujące, oraz AGNES - algorytm hierarchiczny.

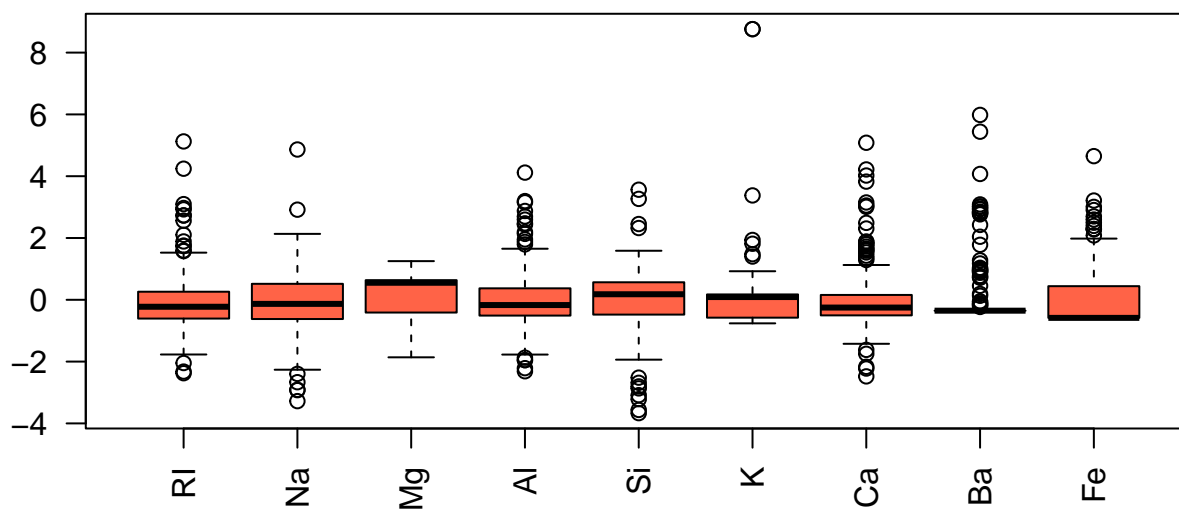
### 2.1 Zbiór danych Glass

Chujowy



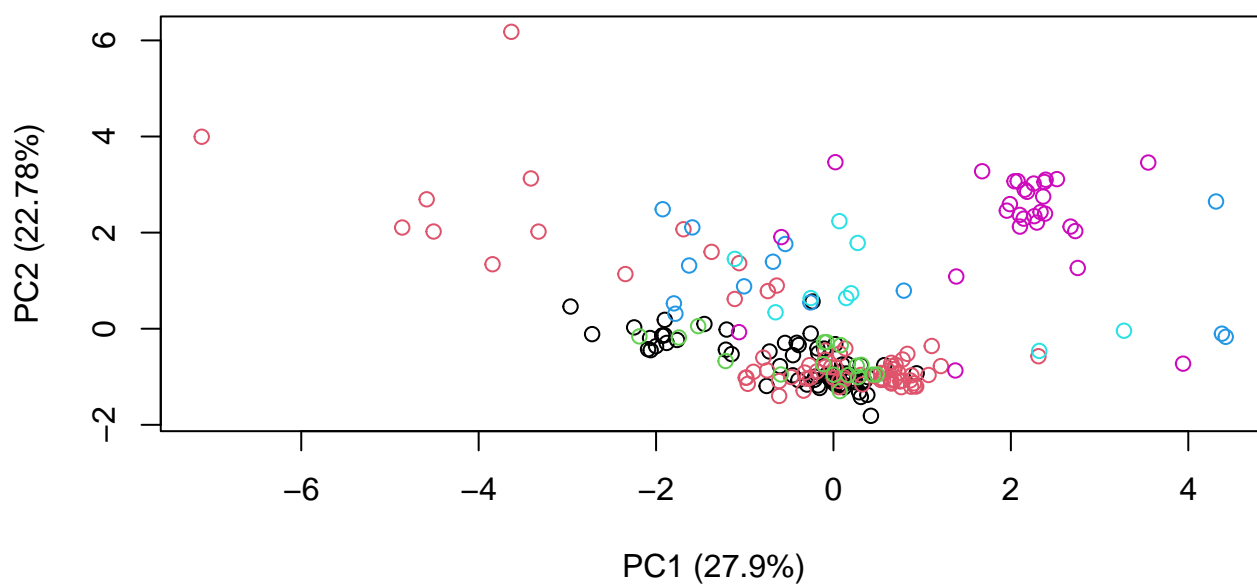
Wykres 1: Wykres pudełkowy, zmienne bez standaryzacji

KURWA STANDARYZACJA MACHEN



Wykres 2: Wykres pudełkowe, po standaryzacji

Teraz jest zajebicie



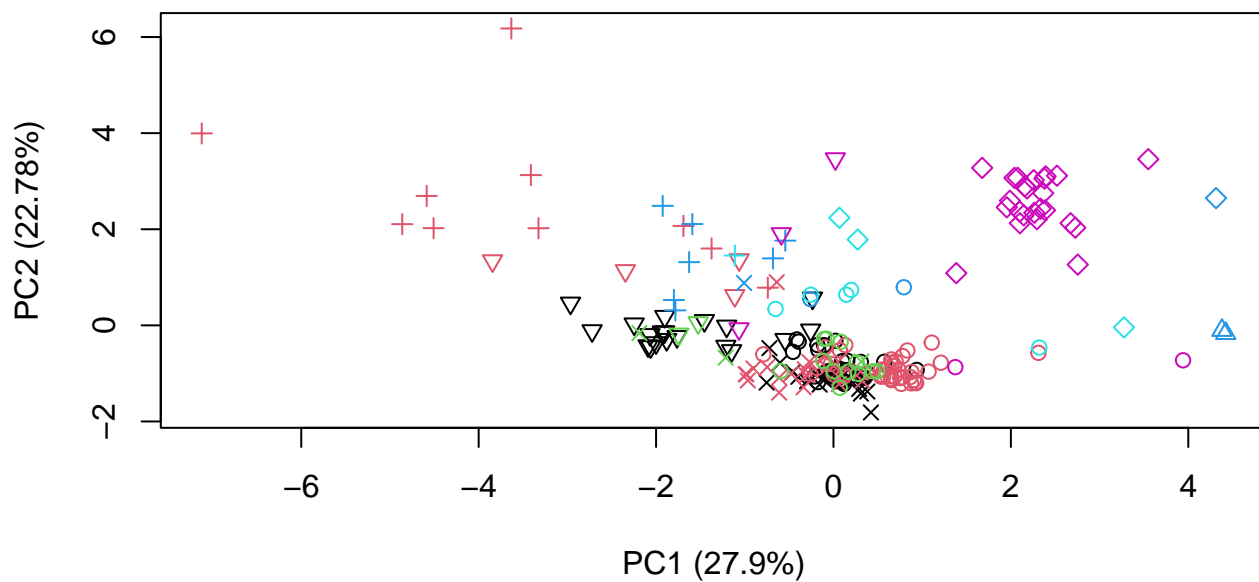
Wykres 3: Wizualizacja danych, PCA

Chuja widać, ciekawe kto wybrał ten zbiór?

## 2.2 Wyniki grupowania

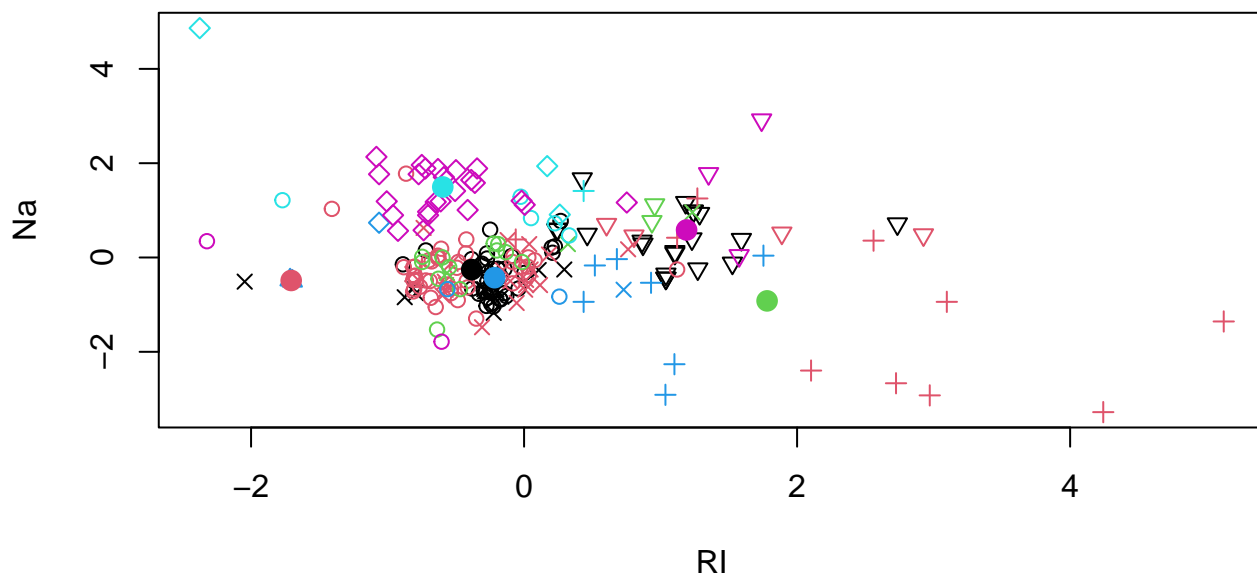
Przeprowadzamy dla rzeczywistej liczby etykiet, która wynosi **6**.

### 2.2.1 k-średnie



Wykres 4: PCA, kolory - rzeczywiste, kształt - wyniki

Gównianie mu poszło



Wykres 5: Wykres RI od Na, aby pokazać gdzie są wyznaczone centra skupień

Centra wywalone w kosmos, ale fajnie

Tabela 23: Macierz błędów; metoda k-średnich

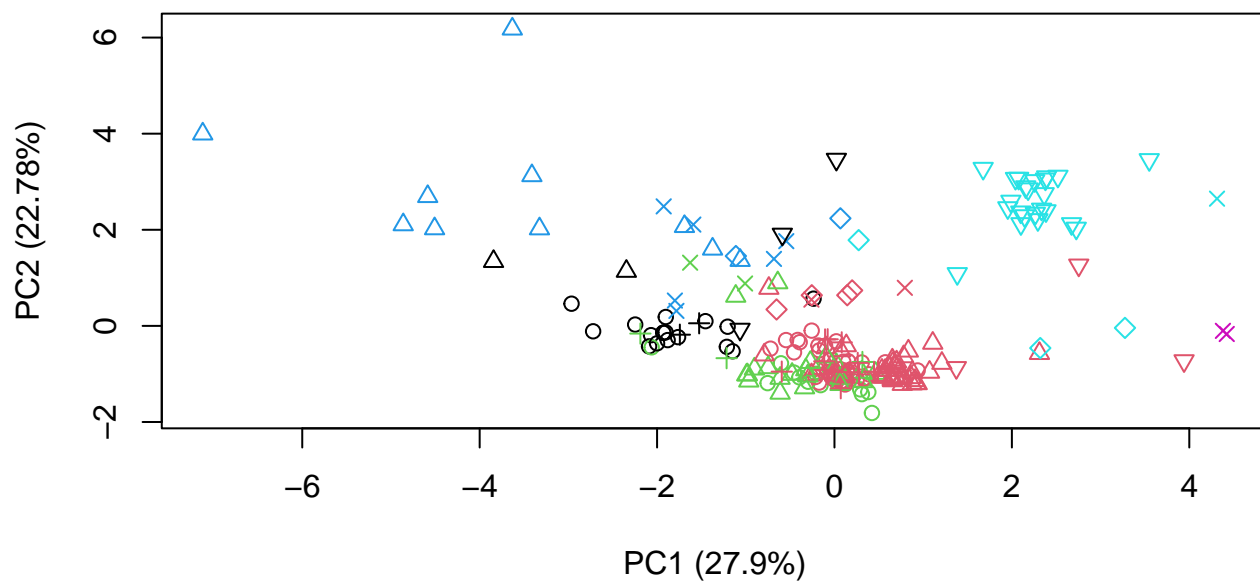
	1	2	3	5	6	7
1	20	4	2	0	0	3

	1	2	3	5	6	7
2	38	42	12	2	5	2
3	12	20	3	1	0	0
5	0	10	0	7	1	0
6	0	0	0	2	0	0
7	0	0	0	1	3	24

Dokładność (macierz): 44.86%.

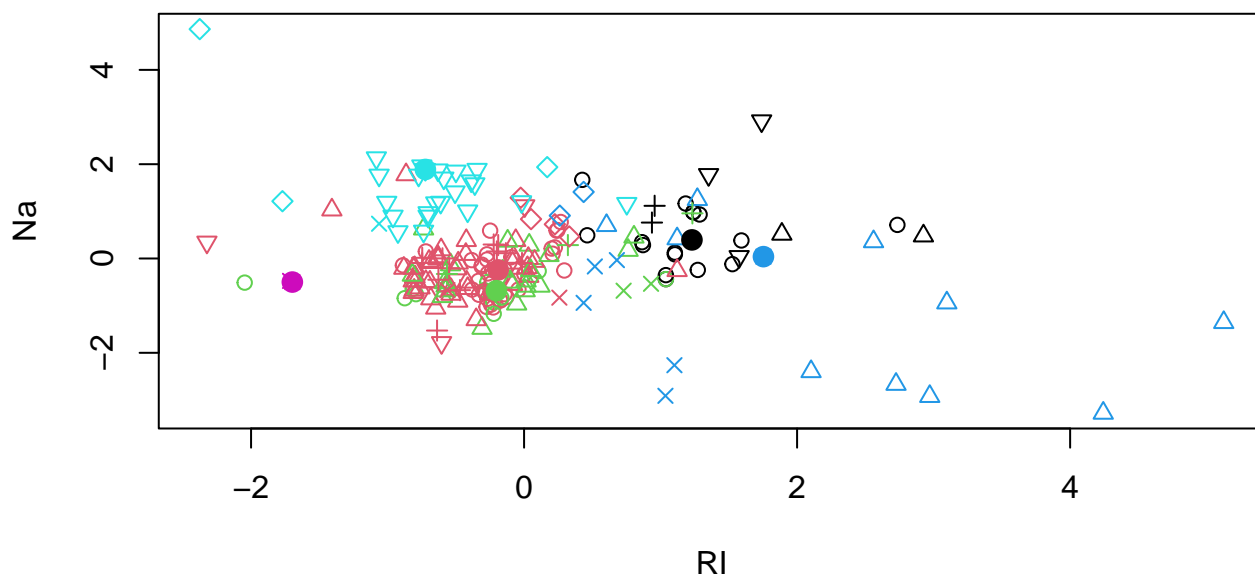
Dokładność (matchClasses, "exact"): 44.86%.

## 2.2.2 Partitioning Around Medoids (PAM)



Wykres 6: coś

Też słabo



Wykres 7: coś, z medoidami

Meh

Tabela 24: Dane medoidów, k=6

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type	PAM
44	1.52210	13.73	3.84	0.72	71.76	0.17	9.74	0.00	0.00	1	1
43	1.51779	13.21	3.39	1.33	72.76	0.59	8.59	0.00	0.00	1	2
33	1.51775	12.85	3.48	1.23	72.97	0.61	8.56	0.09	0.22	1	3
171	1.52369	13.44	0.00	1.58	72.22	0.32	12.24	0.00	0.00	5	4
205	1.51617	14.95	0.00	2.27	73.30	0.00	8.71	0.67	0.00	7	5
173	1.51321	13.00	0.00	3.02	70.70	6.21	6.93	0.00	0.00	5	6

Tabela 25: Macierz błędów; metoda k-średnich

	1	2	3	5	6	7
1	17	2	2	0	0	3
2	40	43	12	2	4	3
3	13	21	3	2	0	0
5	0	10	0	6	2	0
6	0	0	0	1	3	23
7	0	0	0	2	0	0

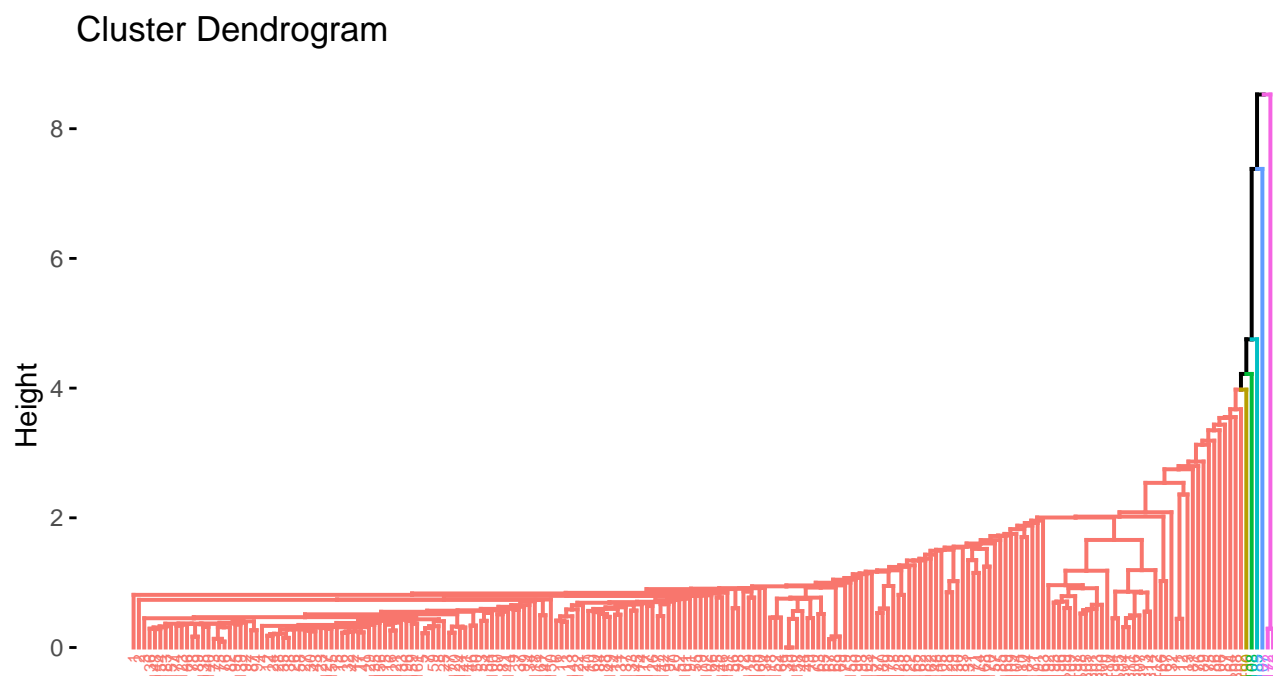
Dokładność: 33.64%.

Dokładność (matchClasses, "exact"): 42.99%.

Gorzej niż k-średnie *shocked emoji*.

## 2.2.3 Agglomerative Nesting (AGNES)

### 2.2.3.1 Najbliższy sąsiad



Wykres 8: AGNES: single linkage

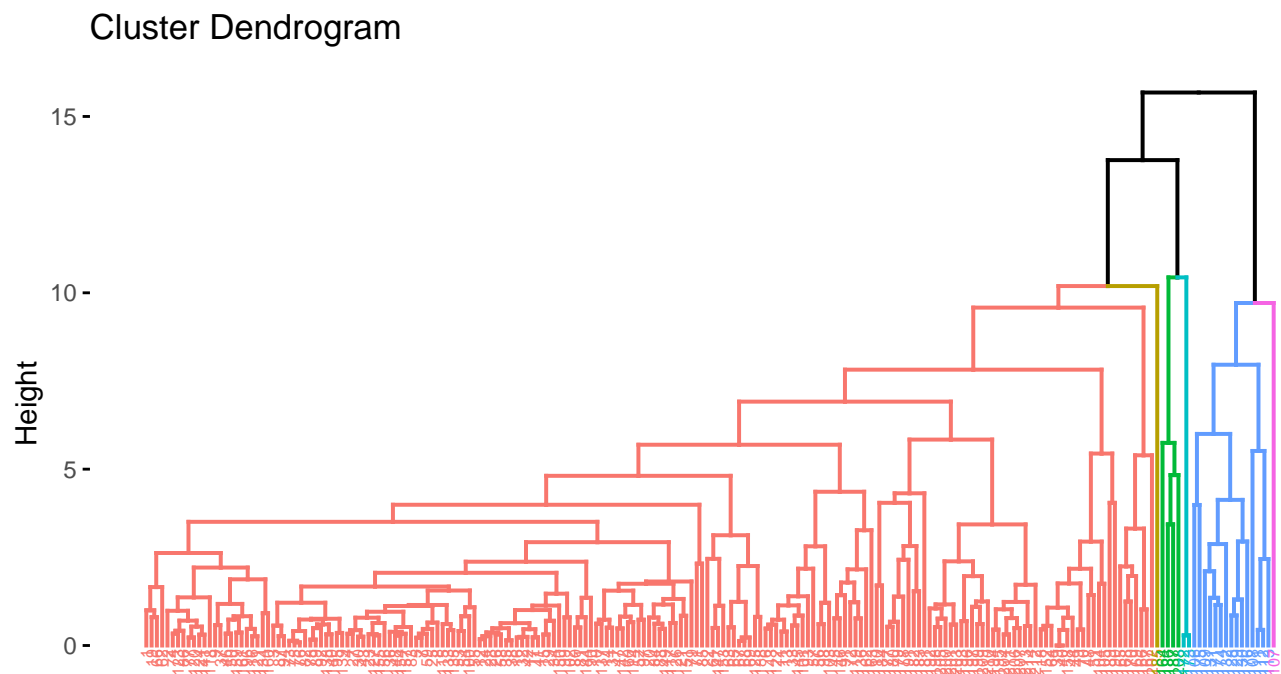
Tabela 26: Macierz błędów; agnes, najbliższy sąsiad

	1	2	3	5	6	7
1	70	74	17	11	8	28
2	0	1	0	0	0	0
3	0	1	0	0	0	0
5	0	0	0	2	0	0
6	0	0	0	0	1	0
7	0	0	0	0	0	1

Dokładność: 35.05%.

Dokładność (matchClasses, “exact”): 36.45%.

#### 2.2.3.2 Najdalszy sąsiad



Wykres 9: AGNES: complete linkage

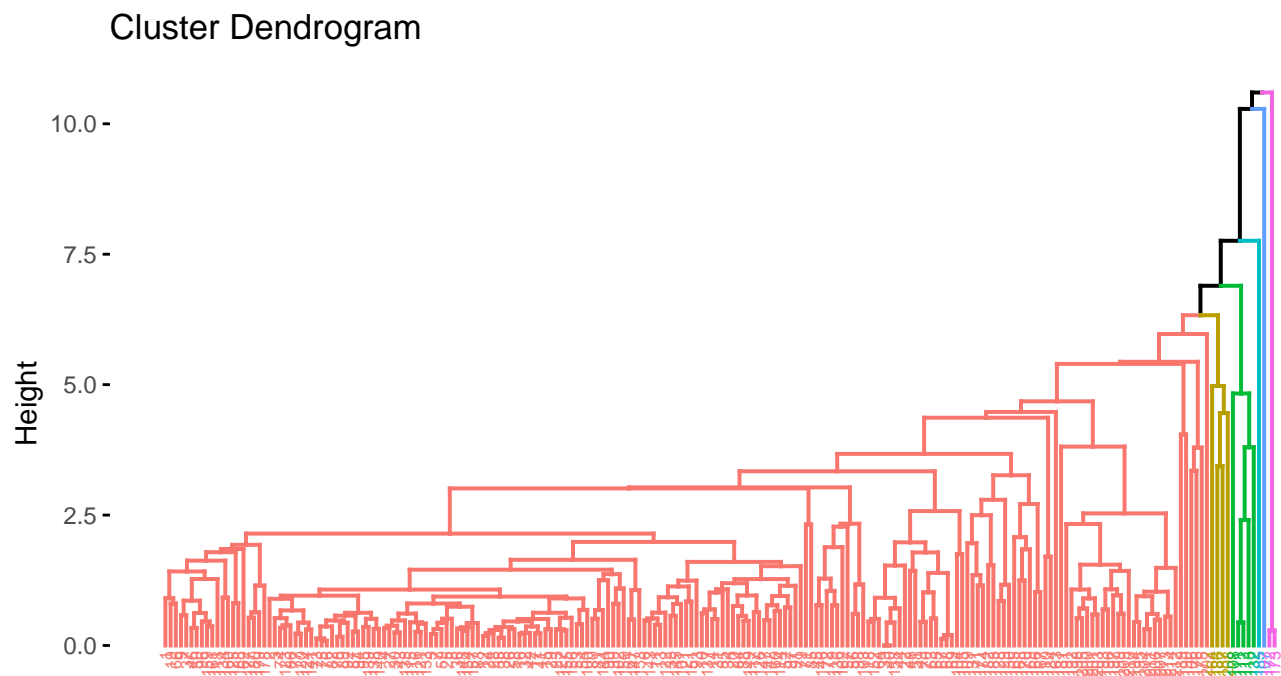
Tabela 27: Macierz błędów; agnes, najdalszy sąsiad

	1	2	3	5	6	7
1	70	64	17	6	8	26
2	0	11	0	4	0	0
3	0	1	0	0	0	0
5	0	0	0	1	0	3
6	0	0	0	2	0	0
7	0	0	0	0	1	0

Dokładność: 38.32%.

Dokładność (matchClasses, “exact”): 40.65%.

#### 2.2.3.3 Średnia odległość



Wykres 10: AGNES: average linkage

Tabela 28: Macierz błędów; agnes, średnia odległość

	1	2	3	5	6	7
1	70	70	17	10	8	26
2	0	1	0	0	0	0
3	0	5	0	0	0	0
5	0	0	0	1	0	3
6	0	0	0	2	0	0
7	0	0	0	0	1	0

Dokładność: 33.64%.

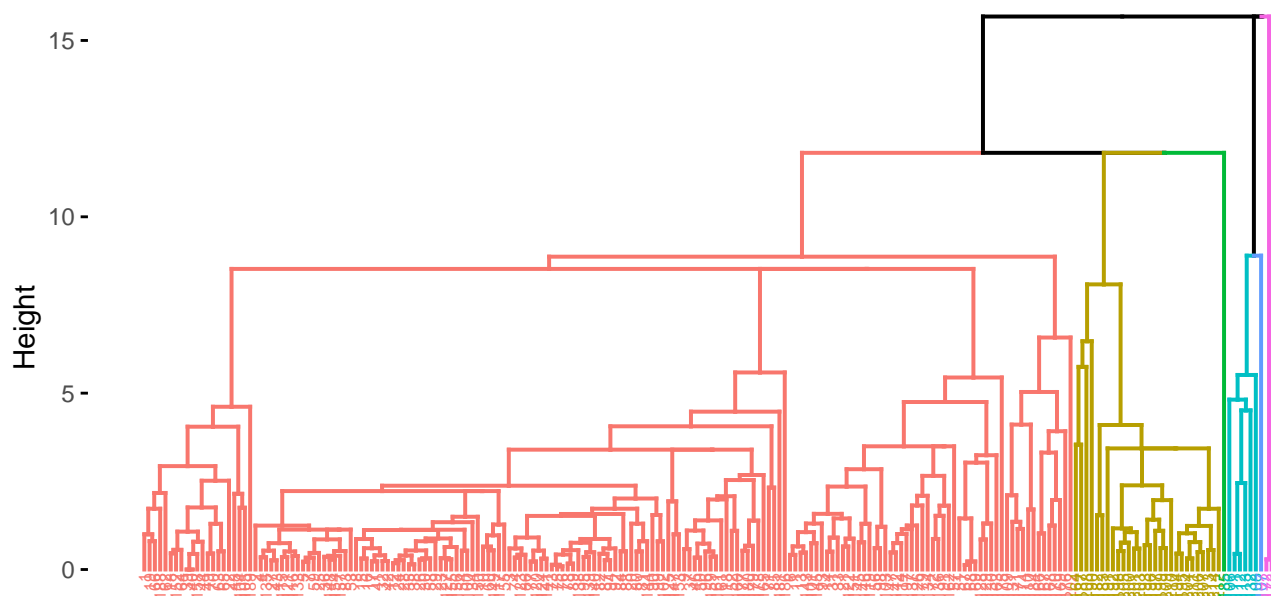
Dokładność (matchClasses, “exact”): 37.85%.

## 2.2.4 Divisive clustering (DIANA)

### 2.2.4.1 Odległość euklidesowa



## Cluster Dendrogram



Wykres 11: AGNES: average linkage

Tabela 29: Macierz błędów; agnes, średnia odległość

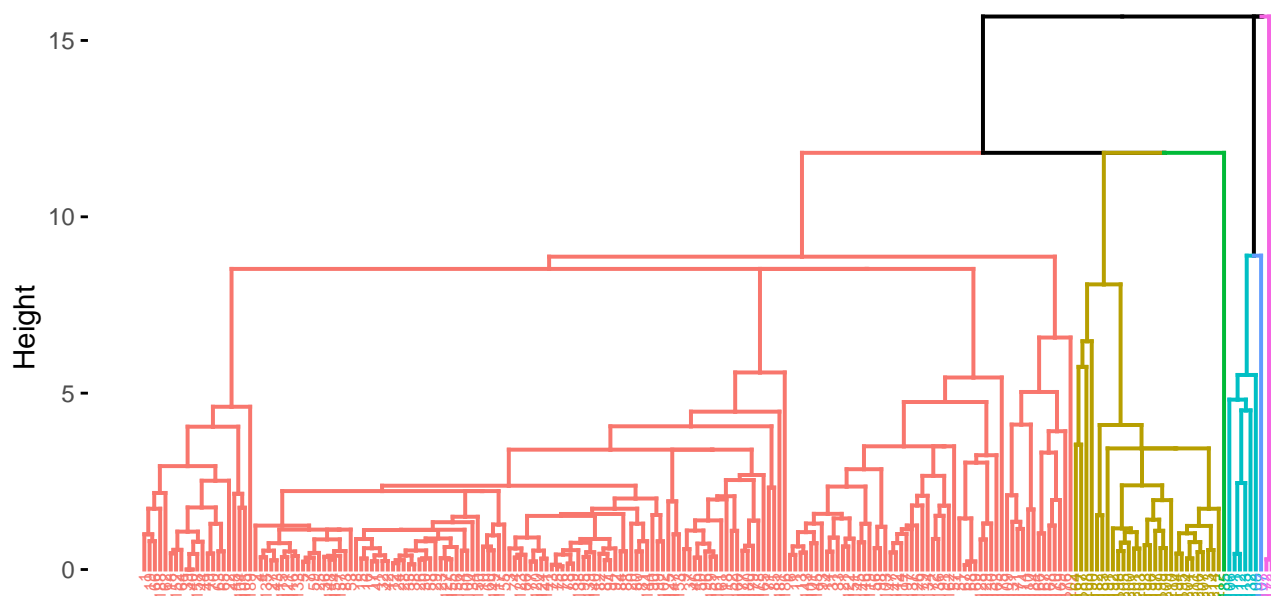
	1	2	3	5	6	7
1	70	69	17	10	6	4
2	0	6	0	0	0	0
3	0	1	0	0	0	0
5	0	0	0	1	2	25
6	0	0	0	2	0	0
7	0	0	0	0	1	0

Dokładność: 35.98%.

Dokładność (matchClasses, “exact”): 48.6%.

### 2.2.4.2 Odległość Manhattan (taksówkowa)

## Cluster Dendrogram



Wykres 12: AGNES: average linkage

Tabela 30: Macierz błędów; agnes, średnia odległość

	1	2	3	5	6	7
1	70	69	17	10	6	4
2	0	6	0	0	0	0
3	0	1	0	0	0	0
5	0	0	0	1	2	25
6	0	0	0	2	0	0
7	0	0	0	0	1	0

Dokładność: 35.98%.

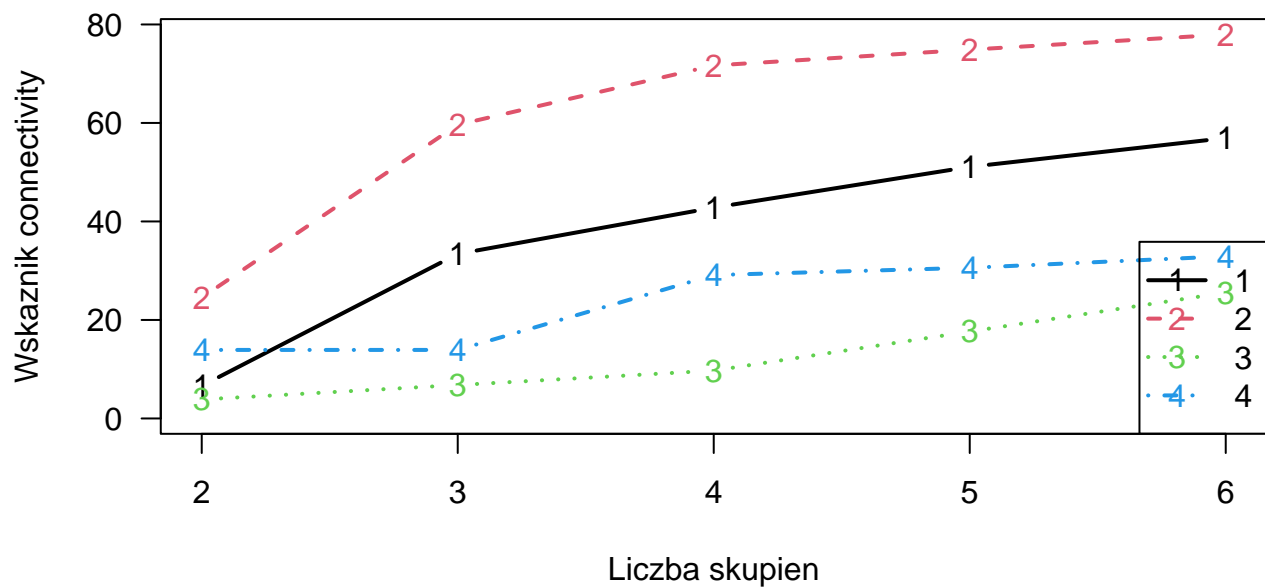
Dokładność (matchClasses, "exact"): 48.6%.

## 2.3 Ocena jakości grupowania i wizualizacja najlepszych wyników

### 2.3.1 Ocena

Legenda:

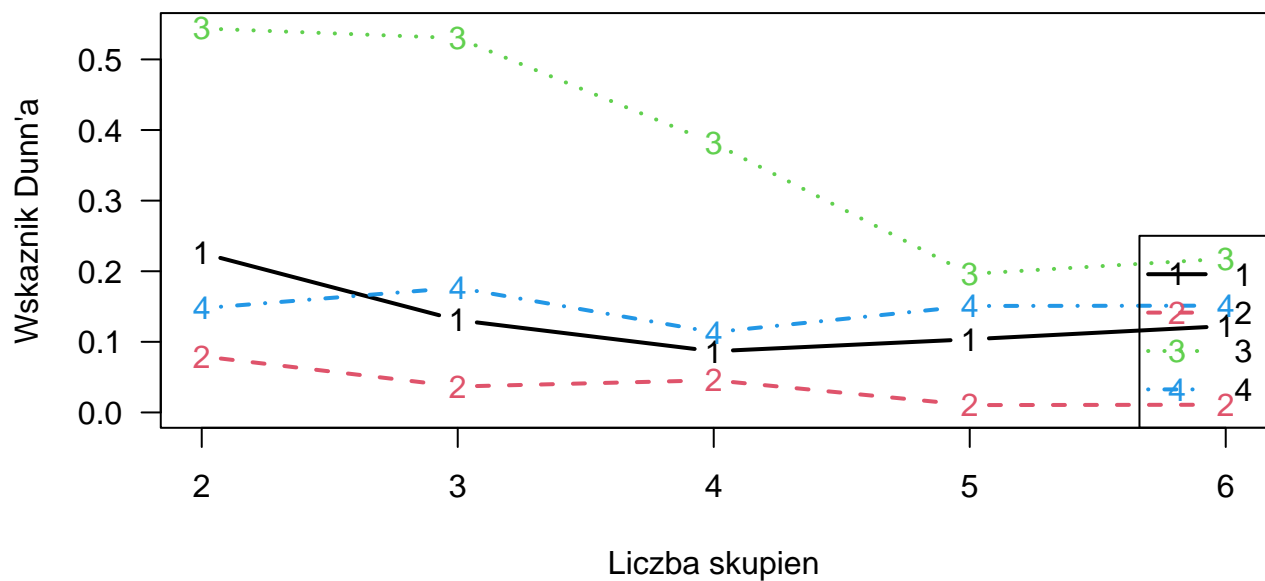
- 1 - kmeans,
- 2 - PAM,
- 3 - AGNES,
- 4 - DIANA.



Wykres 13: Connectivity

Connectivity - im mniejszy, tym lepszy:

- kmeans - 2,
- PAM - 2,
- AGNES - 2 (najlepszy),
- DIANA - 3.

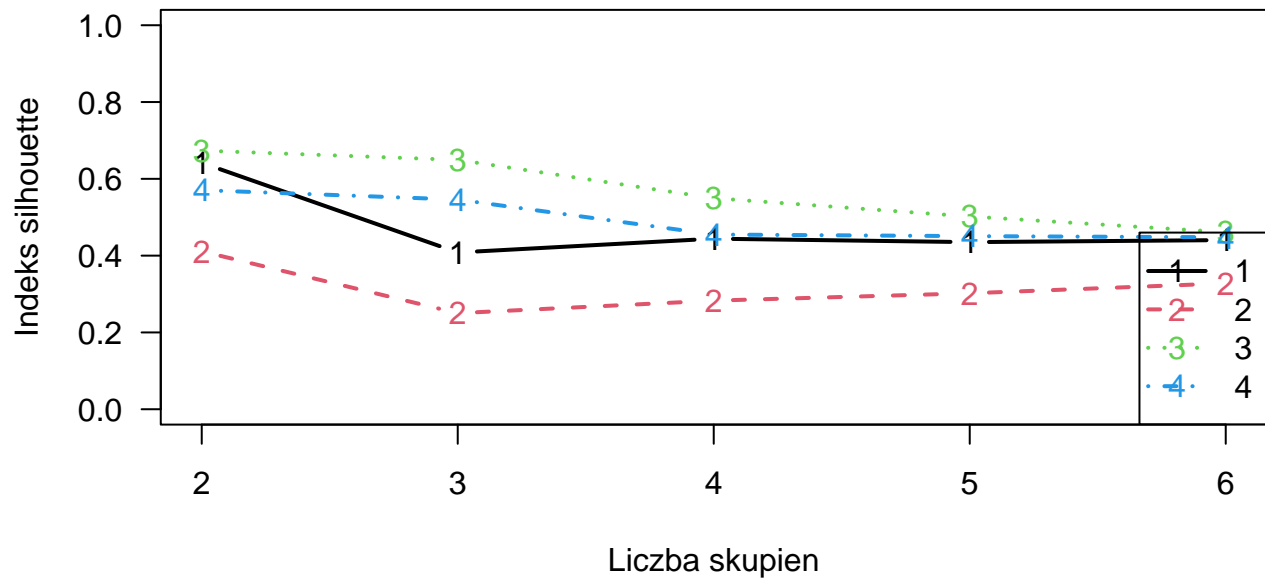


Wykres 14: Dunn

Dunn - im większy, tym lepszy:

- kmeans - 2,
- PAM - 2,
- AGNES - 2 (najlepszy),

- DIANA - 3.



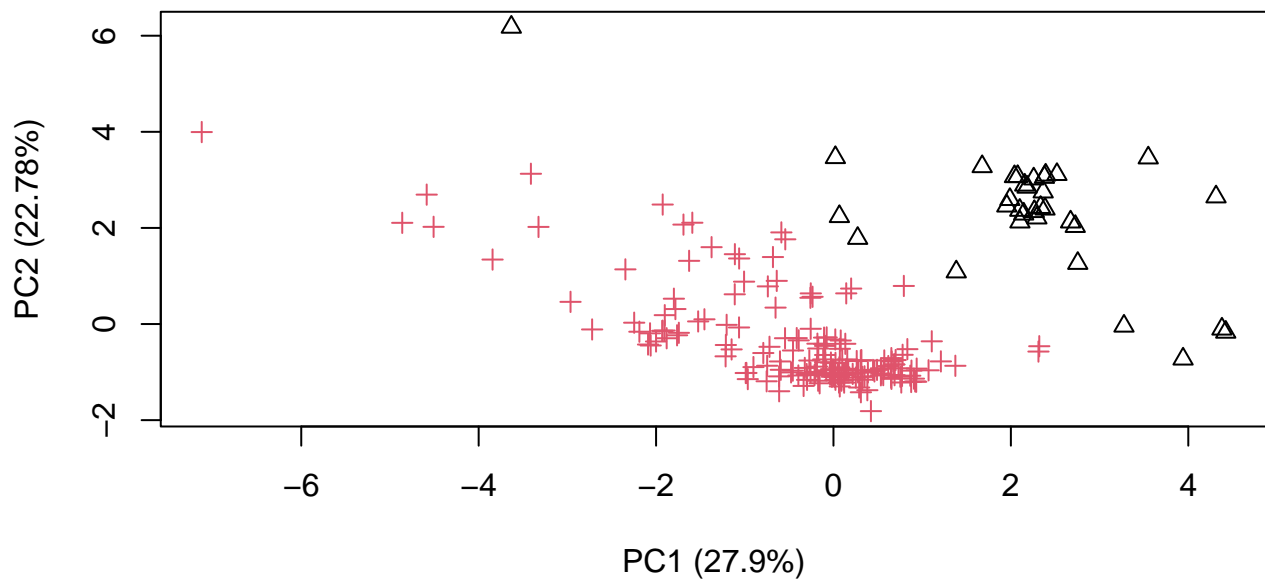
Wykres 15: Silhouette

Dunn - im większy, tym lepszy:

- kmeans - 2,
- PAM - 2,
- AGNES - 2 (najlepszy),
- DIANA - 2.

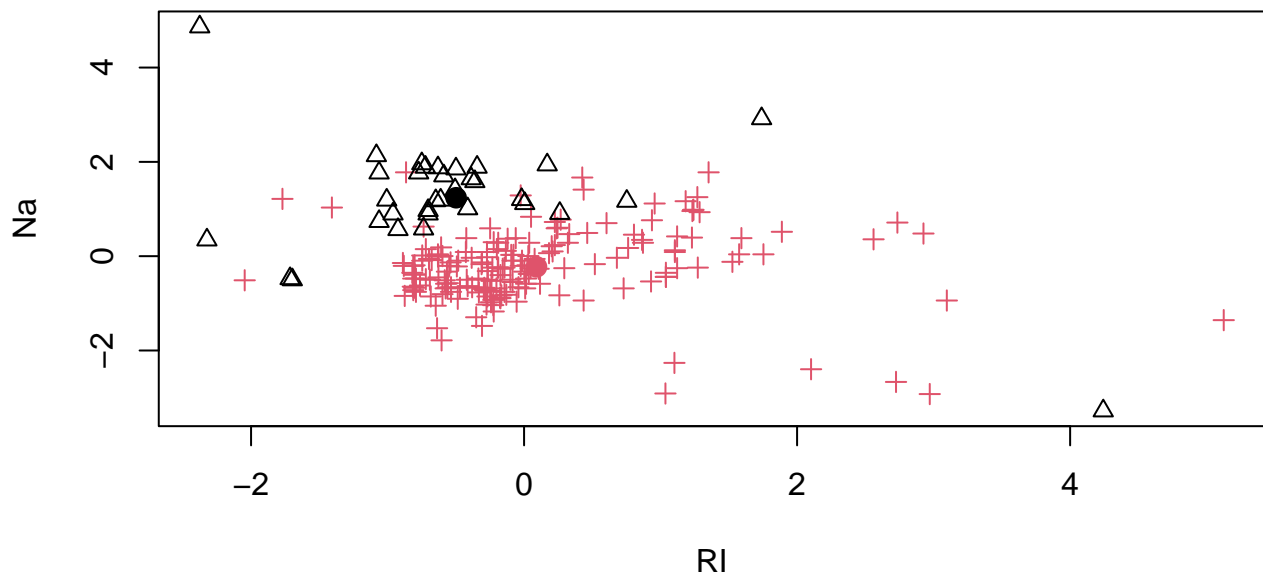
### 2.3.2 Wizualizacja

#### 2.3.2.1 k-średnie



Wykres 16: PCA, kolory - rzeczywiste, kształt - wyniki, k=2

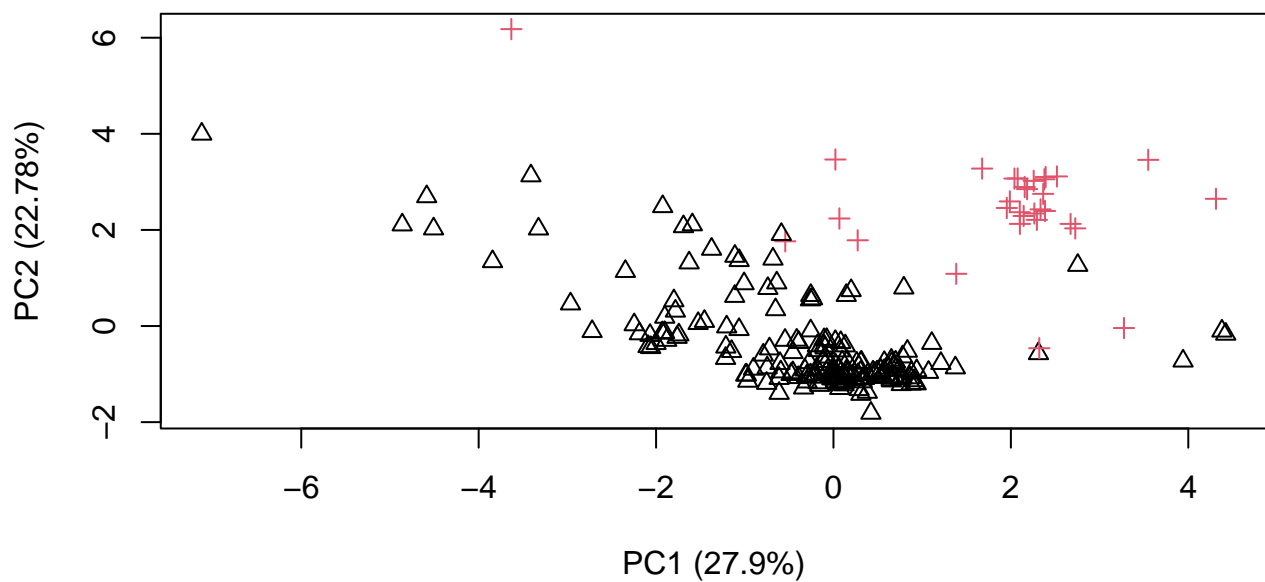
Gównianie mu poszło



Wykres 17: Wykres RI od Na, aby pokazać gdzie są wyznaczone centra skupień, k=2

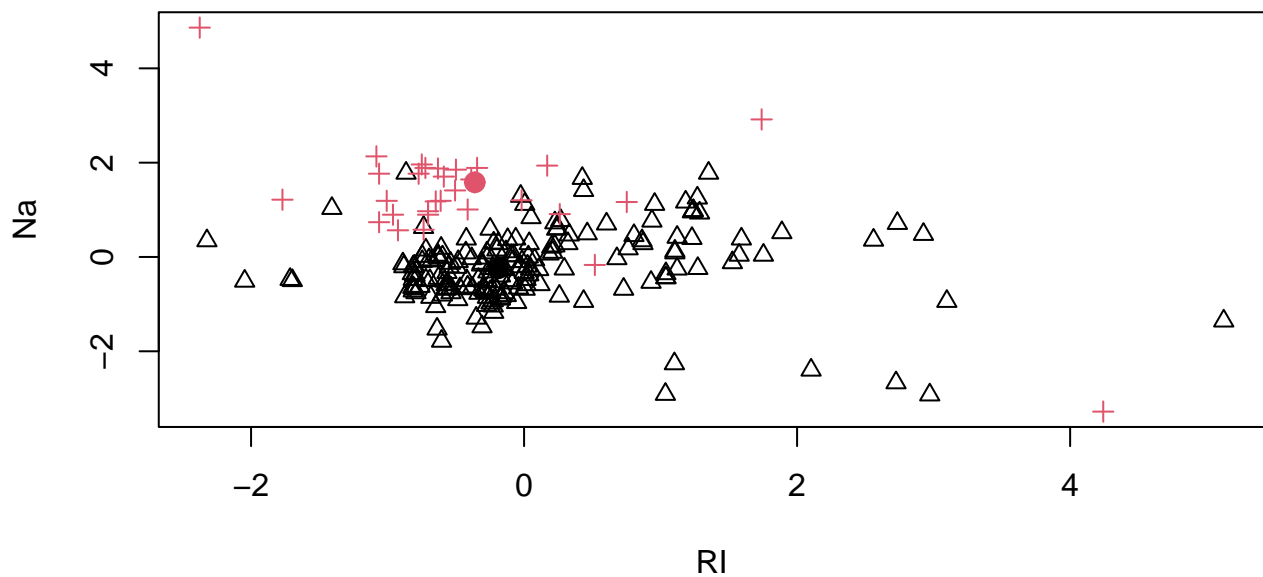
Centra wywalone w kosmos, ale fajnie  
Dokładność (matchedClasses): 96.19%.

### 2.3.2.2 PAM



Wykres 18: coś, k=2

Też słabo



Wykres 19: coś, z medoidami, k=2

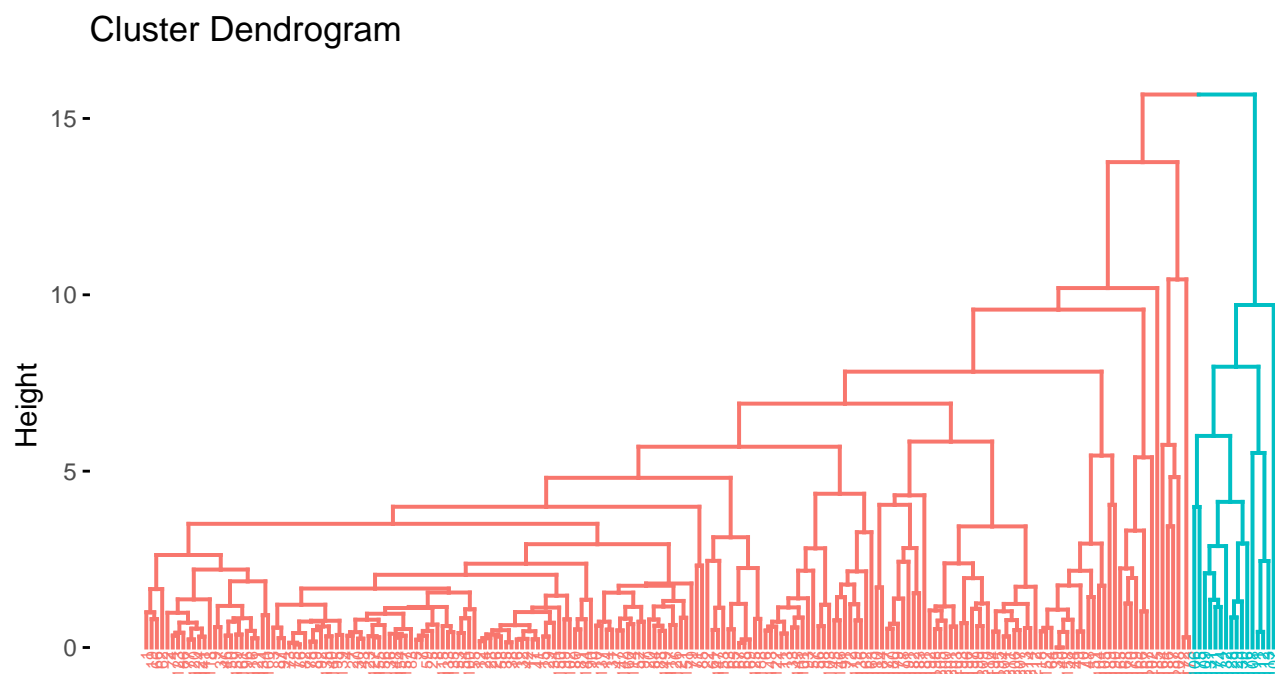
Meh

Dokładność (matchedClasses): 94.29%.

Tabela 31: Dane medoidów, k=2

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type	PAM
43	1.51779	13.21	3.39	1.33	72.76	0.59	8.59	0.00	0	1	1
198	1.51727	14.70	0.00	2.34	73.28	0.00	8.95	0.66	0	7	2

### 2.3.2.3 AGNES



Wykres 20: AGNES: complete linkage,  $k=2$

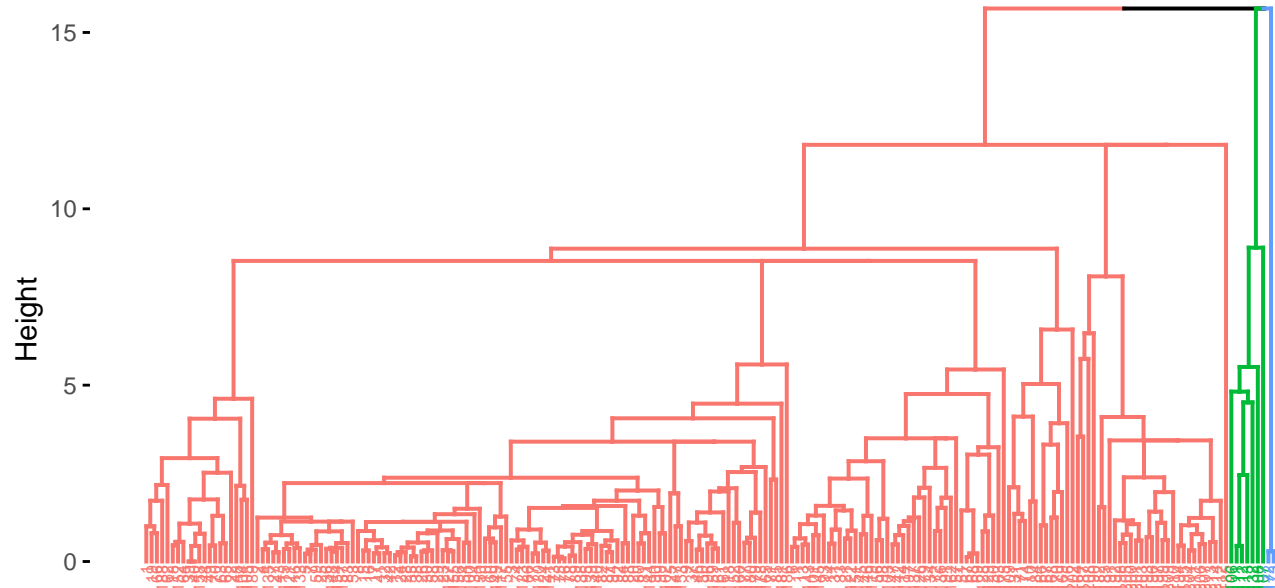
Tylko dla najdalszych sąsiadów, bo dał najlepsze wyniki.

Dokładność (matchedClasses): 87.64%.

#### 2.3.2.4 DIANA



## Cluster Dendrogram



Wykres 21: DIANA,  $k=3$

Dokładność (matchedClasses): 49.69%.

## 2.4 Wnioski

## 3 Podsumowanie

PS. Czas wykonywania kodu wynosi 7 minut i 3 sekund.