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

Filip Michewicz 282239
Wiktor Niedźwiedzki 258882

18 czerwca 2025 Anno Domini

Spis treści

Spis wykresów

Spis tabel

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).

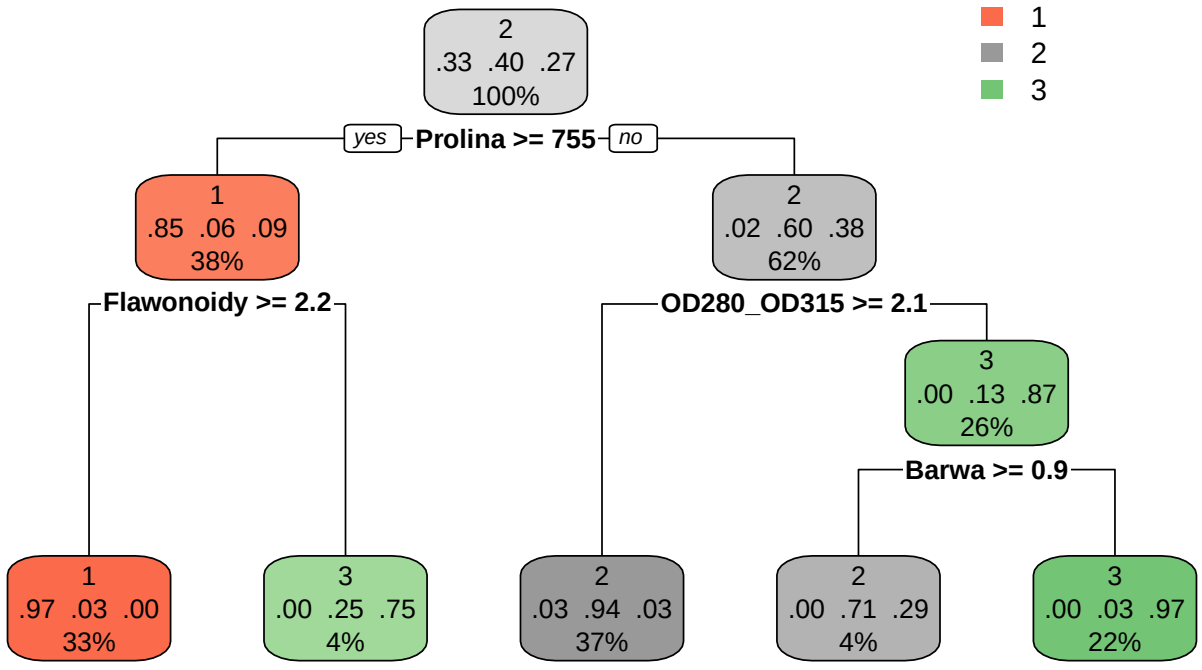
Zadanie zostanie wykonane na zbiorze danych *wine*, którego szczegółowy opis znajduje się w poprzednim raporcie.

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).

Na wykresie przedstawiono pojedyńcze drzewo klasyfikacyjne.



Wykres 1: Pojedyńcze drzewo klasyfikacyjne

Błąd estymowany metodą bootstrap .632+ wynosi 9.1%. Na jego podstawie określimy poprawę modelu po zastosowaniu metod uczenia zespołowego.

W tej części analizy do oceny wydajności modelu zostanie wykorzystana wyłącznie metoda .632+, ponieważ koryguje ona obciążenie estymatora (bias) i dostarcza bardziej wiarygodnej oceny błędu generalizacji, zwłaszcza przy ryzyku przeuczenia i ograniczonej liczbie próbek.

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	41.52	25.72	42.43	63.67	56.94	63.24	65.60	65.03
Random Forest	87.65	83.70	86.83	89.45	82.73	88.72	82.30	84.78
Boosting	72.95	67.31	67.28	73.75	63.21	67.37	70.96	71.30

1.2 Metoda wektorów nośnych (SVM)

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

Metoda SVM jest jedną z najczęściej stosowanych technik uczenia maszynowego w zadaniach klasyfikacyjnych. Jej podstawowym celem jest wyznaczenie hiperpłaszczyzny maksymalnie oddzielającej obserwacje należące do różnych klas, przy jednoczesnym maksymalizowaniu marginesu między klasami.

Dzięki zastosowaniu funkcji jądrowych (kernel functions), SVM umożliwia również skuteczną klasyfikację danych nieliniowo separowalnych poprzez odwzorowanie ich do przestrzeni o wyższej liczbie wymiarów. W niniejszej analizie zostaną porównane różne funkcje jądrowe, w tym liniowa, wielomianowa oraz radialna (RBF), w kontekście ich wpływu na jakość klasyfikacji.

1.2.1 Jądro liniowe

Przeanalizowano skuteczność klasyfikacyjną z zastosowaniem **jądra liniowego**, zarówno **bez skalowania danych**, jak i **po ich skalowaniu**.

Porównanie wyników dla obu wariantów (ze skalowaniem i bez) pozwala ocenić wpływ przeskalowania zmiennych na jakość klasyfikacji. Ponieważ SVM opiera się na obliczeniach odległości i iloczynów skalarnych, skalowanie danych może znacząco wpłynąć na działanie algorytmu, szczególnie gdy zmienne wejściowe różnią się skalą lub jednostką.

Tabela 2: Jądro liniowe - bez skalowania

	0.001	0.01	0.1	1	10	100	1000
Wielokrotny podział	40.00	98.17	97.00	95.50	96.50	96.33	95.67
Cross-validation	45.08	57.14	92.88	93.63	93.63	93.63	93.63
Bootstrap	37.37	95.86	97.21	97.47	95.47	96.75	96.23
Średnio	40.82	83.72	95.70	95.53	95.20	95.57	95.18

Tabela 3: Jądro liniowe - ze skalowaniem

	0.001	0.01	0.1	1	10	100	1000
Wielokrotny podział	38.50	97.50	97.17	96.50	96.00	96.50	96.00
Cross-validation	32.71	50.63	93.26	93.07	93.07	93.07	93.07
Bootstrap	38.65	96.34	97.28	95.46	97.25	96.45	96.80
Średnio	36.62	81.49	95.90	95.01	95.44	95.34	95.29

Z analizy Tabeli 2. oraz Tabeli 3. wynika, że optymalną wartością parametru **C** w klasyfikacji SVM z jądrem liniowym jest **0.1**. Dla tej wartości obserwuje się najwyższą średnią skuteczność klasyfikacji zarówno bez skalowania, jak i po skalowaniu danych. Ponadto, zastosowanie skalowania cech nieznacznie poprawia wyniki, co wskazuje na korzystny wpływ normalizacji na efektywność modelu.

Wyższe wartości parametru **C** nie przekładają się na istotną poprawę skuteczności klasyfikacji, a w niektórych przypadkach powodują nawet jej nieznaczny spadek. Wynika to z faktu, że zbyt duża wartość **C** powoduje nadmierne dopasowanie modelu do danych treningowych (overfitting). W konsekwencji, mimo że model stara się minimalizować błędy na zbiorze treningowym, jego efektywność na danych testowych nie ulega poprawie, co potwierdzają uzyskane wyniki.

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	36.33	93.17	96.17	95.00	95.00	95.50	94.33	94.00	95.83	95.83
0.01	40.00	97.00	94.67	95.83	94.50	95.17	93.67	95.17	95.33	96.33
0.1	39.50	95.17	92.50	95.50	95.33	95.00	96.33	94.83	94.50	96.00
1	39.17	94.67	97.33	94.17	95.17	95.83	95.67	95.33	94.67	95.67
10	39.83	95.67	94.83	94.50	94.33	95.17	96.00	94.83	96.00	96.00
100	83.33	95.83	96.17	96.17	96.50	95.67	96.83	96.00	94.83	94.33
1000	95.83	95.67	96.00	95.17	96.00	95.00	95.33	94.33	93.67	95.67

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	35.33	95.33	96.17	94.17	96.33	94.50	96.00	96.17	96.50	96.33
0.01	36.33	96.50	95.33	95.33	95.17	94.67	95.67	94.17	95.67	95.50
0.1	37.33	96.50	95.00	95.50	94.83	96.17	94.83	96.00	95.67	95.50
1	38.83	96.17	95.67	95.67	95.17	96.17	95.33	96.50	95.00	95.33
10	40.67	96.00	95.83	96.00	95.33	96.33	95.83	95.17	95.50	95.17
100	88.00	95.17	94.67	94.83	95.17	95.17	94.67	95.67	94.33	96.17
1000	95.33	94.00	94.50	95.00	96.00	96.00	94.00	94.17	95.83	95.00

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	94.93	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35
0.01	39.80	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35
0.1	39.80	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35
1	39.80	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35
10	44.28	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35
100	88.69	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35
1000	94.93	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35	94.35

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	39.80	95.52	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
0.01	39.80	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
0.1	39.80	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
1	39.80	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
10	43.73	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
100	87.06	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67	96.67
1000	96.63	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	39.17	93.99	94.70	94.28	94.58	94.08	94.64	94.61	93.44	93.52
0.01	34.68	94.06	94.91	92.90	94.17	95.14	93.69	94.03	92.45	93.81
0.1	36.34	95.98	94.19	93.71	93.74	93.68	94.77	94.65	93.36	95.36
1	39.45	95.53	94.11	93.63	93.24	94.75	95.42	92.83	93.78	93.91
10	41.55	93.83	93.23	94.61	95.77	92.71	93.80	94.85	95.23	94.31
100	87.27	94.04	93.04	94.96	93.93	93.38	95.03	95.09	94.76	95.61
1000	94.04	95.09	93.85	93.40	94.63	93.78	93.97	95.72	93.37	94.56

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	37.99	92.32	92.90	95.57	94.85	93.88	95.23	94.54	91.80	95.98
0.01	37.83	92.84	92.76	92.69	94.38	94.35	93.98	92.91	94.11	93.42
0.1	38.16	94.86	93.58	94.89	94.22	94.03	94.30	94.85	94.28	95.97
1	35.54	94.72	94.18	94.05	94.75	95.03	93.85	95.29	94.41	95.22
10	40.44	95.11	93.75	93.20	93.50	94.74	93.99	93.61	93.87	93.27
100	89.20	95.34	92.66	95.52	94.58	95.62	95.79	92.88	95.09	94.60
1000	94.43	94.54	94.60	94.56	94.73	95.03	94.82	94.99	95.17	96.20

Najlepsza gamma: **7.78**, najlepsza kara: **1**. 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	86.83	95.83	87.83	87.83	76.5	81.67

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	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84
0.01	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84
0.1	79.80	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84
1	98.89	57.75	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84
10	97.16	61.67	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84
100	96.05	61.67	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84
1000	96.05	61.67	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84

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	39.90	39.90	39.90	39.9	39.9	39.9	39.9	39.9	39.9	39.9
0.01	39.90	39.90	39.90	39.9	39.9	39.9	39.9	39.9	39.9	39.9
0.1	82.58	39.90	39.90	39.9	39.9	39.9	39.9	39.9	39.9	39.9
1	98.30	58.95	39.90	39.9	39.9	39.9	39.9	39.9	39.9	39.9
10	97.78	62.32	40.46	39.9	39.9	39.9	39.9	39.9	39.9	39.9
100	96.63	62.32	40.46	39.9	39.9	39.9	39.9	39.9	39.9	39.9
1000	96.63	62.32	40.46	39.9	39.9	39.9	39.9	39.9	39.9	39.9

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	39.87	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	82.55	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
1	98.33	56.21	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
10	98.30	61.34	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
100	97.75	61.34	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87
1000	97.75	61.34	39.87	39.87	39.87	39.87	39.87	39.87	39.87	39.87

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	40.00	40.00	40	40	40	40	40	40	40	40
0.01	40.00	40.00	40	40	40	40	40	40	40	40
0.1	79.25	40.00	40	40	40	40	40	40	40	40
1	98.89	57.91	40	40	40	40	40	40	40	40
10	97.22	61.27	40	40	40	40	40	40	40	40
100	96.67	61.27	40	40	40	40	40	40	40	40
1000	96.67	61.27	40	40	40	40	40	40	40	40

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	40.82	37.68	34.33	38.18	39.02	36.42	37.85	34.30	34.92	35.34
0.01	33.43	35.49	37.56	35.96	38.23	34.65	39.08	37.99	38.62	36.47
0.1	63.12	36.88	36.58	38.88	36.32	36.84	37.49	37.42	37.43	39.90
1	96.70	46.32	38.24	39.32	33.72	38.79	39.41	38.04	38.75	36.06
10	95.99	54.32	40.76	37.48	36.56	39.59	36.49	37.09	39.41	38.02
100	96.37	50.12	39.14	40.81	36.05	37.33	38.77	36.47	36.03	36.45
1000	97.44	52.29	40.78	36.25	37.65	36.76	39.14	38.06	36.61	39.31

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	39.98	30.30	39.33	34.66	35.77	36.13	36.99	36.33	36.23	36.26
0.01	38.18	33.50	38.61	36.13	36.61	35.86	38.84	36.60	37.62	38.94
0.1	56.80	38.34	37.90	37.40	40.68	38.13	36.95	40.13	36.05	38.94
1	96.83	47.19	35.87	41.13	36.81	35.02	38.19	39.70	37.36	39.01
10	96.75	54.60	41.25	41.18	35.95	38.43	40.43	37.50	38.40	36.81
100	96.69	51.39	35.81	37.77	32.92	35.75	38.01	36.43	37.72	39.56
1000	97.60	54.58	36.00	38.14	40.06	37.82	38.28	39.85	38.00	37.31

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	36.83	31.83	37.33	39.83	41.83	40.00	36.17	39.50	38.83	36.50
0.01	39.33	60.83	57.17	66.67	57.67	57.00	59.33	56.83	63.33	62.83
0.1	41.33	87.67	87.33	89.33	86.33	88.67	88.00	85.67	89.33	86.67
1	97.50	83.00	82.83	80.83	81.33	80.83	81.00	81.67	82.50	79.83
10	98.17	82.00	80.17	81.17	82.00	81.67	83.17	79.00	80.83	81.33
100	96.33	81.33	80.17	80.33	82.17	80.17	81.33	77.67	82.33	78.00
1000	96.17	84.00	79.50	81.83	80.83	80.67	82.33	81.67	77.83	79.50

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	39.83	40.00	37.00	38.50	44.17	43.50	38.50	43.00	38.00	38.33
0.01	36.67	54.83	65.00	56.33	60.83	63.00	65.00	60.50	57.50	58.67
0.1	37.67	86.83	87.67	87.50	85.83	87.33	87.00	84.67	87.00	87.83
1	97.33	83.17	83.67	80.00	81.17	82.33	81.00	82.83	83.33	80.17
10	97.83	83.50	81.67	81.00	79.50	78.50	81.83	81.67	81.00	81.50
100	96.33	81.17	81.33	81.00	79.33	80.50	82.67	77.33	82.17	82.33
1000	96.50	84.17	80.17	80.50	80.50	82.83	81.83	81.33	80.00	82.50

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	39.90	39.90	39.90	39.90	39.90	39.90	39.90	39.90	39.90	39.90
0.01	39.90	78.59	78.59	79.15	78.59	79.15	79.15	79.15	79.15	79.15
0.1	42.71	87.09	86.57	85.42	86.54	86.54	87.09	87.09	87.09	85.98
1	97.75	79.74	79.71	80.85	77.45	75.75	75.85	73.56	74.74	75.85
10	97.75	80.88	80.36	81.99	78.59	78.04	76.37	75.26	74.71	75.85
100	95.46	82.55	80.95	80.29	79.74	80.29	77.52	76.44	75.88	76.44
1000	96.05	79.22	80.95	81.41	79.74	81.96	79.74	76.41	76.47	78.14

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	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84	39.84
0.01	39.84	82.68	83.27	83.27	83.27	83.27	83.27	83.27	83.27	83.27
0.1	42.09	87.61	85.95	85.36	85.92	85.36	85.92	86.47	86.47	86.47
1	98.89	82.03	77.55	76.96	76.41	77.52	73.07	75.33	74.71	74.71
10	97.16	79.80	81.44	78.07	79.22	77.55	78.10	79.25	78.07	78.07
100	95.46	79.77	83.10	79.77	76.99	77.52	77.55	79.25	78.66	78.66
1000	96.01	78.66	83.66	79.77	79.80	78.10	75.29	79.22	77.52	78.10

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	39.28	35.26	35.19	38.34	37.50	40.17	36.51	35.34	36.49	37.89
0.01	39.08	65.94	63.59	57.18	61.98	58.70	55.16	63.59	61.90	64.92
0.1	36.36	87.56	84.88	86.91	83.69	87.60	84.29	86.37	84.18	83.47
1	97.26	79.91	80.61	82.12	80.90	82.45	81.10	81.12	78.30	78.86
10	96.40	80.51	79.50	77.76	80.29	79.04	81.37	84.14	78.88	80.50
100	96.24	79.28	78.46	81.51	79.92	79.94	79.04	79.34	79.69	80.90
1000	96.39	82.88	77.45	80.09	81.36	78.23	80.24	81.02	82.08	80.70

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	38.35	35.36	40.41	34.11	35.09	38.37	32.98	36.73	36.71	36.73
0.01	36.35	64.19	57.76	59.33	62.70	58.73	62.30	61.89	63.67	57.05
0.1	34.90	85.27	86.89	86.38	84.98	86.02	83.63	85.47	86.99	84.66
1	97.61	80.99	79.75	80.34	80.10	82.19	80.91	78.29	80.93	83.12
10	97.37	77.79	81.62	81.61	79.83	80.22	78.69	83.38	79.43	78.62
100	96.85	77.92	80.59	79.81	77.49	77.49	77.22	80.88	77.63	80.57
1000	96.07	78.06	79.02	79.51	82.15	82.44	77.72	78.57	82.14	80.14

elo

1.3 Wnioski

e

2 Analiza skupień – algorytmy grupujące i hierarchiczne

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

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

To zadanie zostanie wykonane już na innym danych, którymi będzie zbiór *glass*.

2.1 Charakterystyka danych

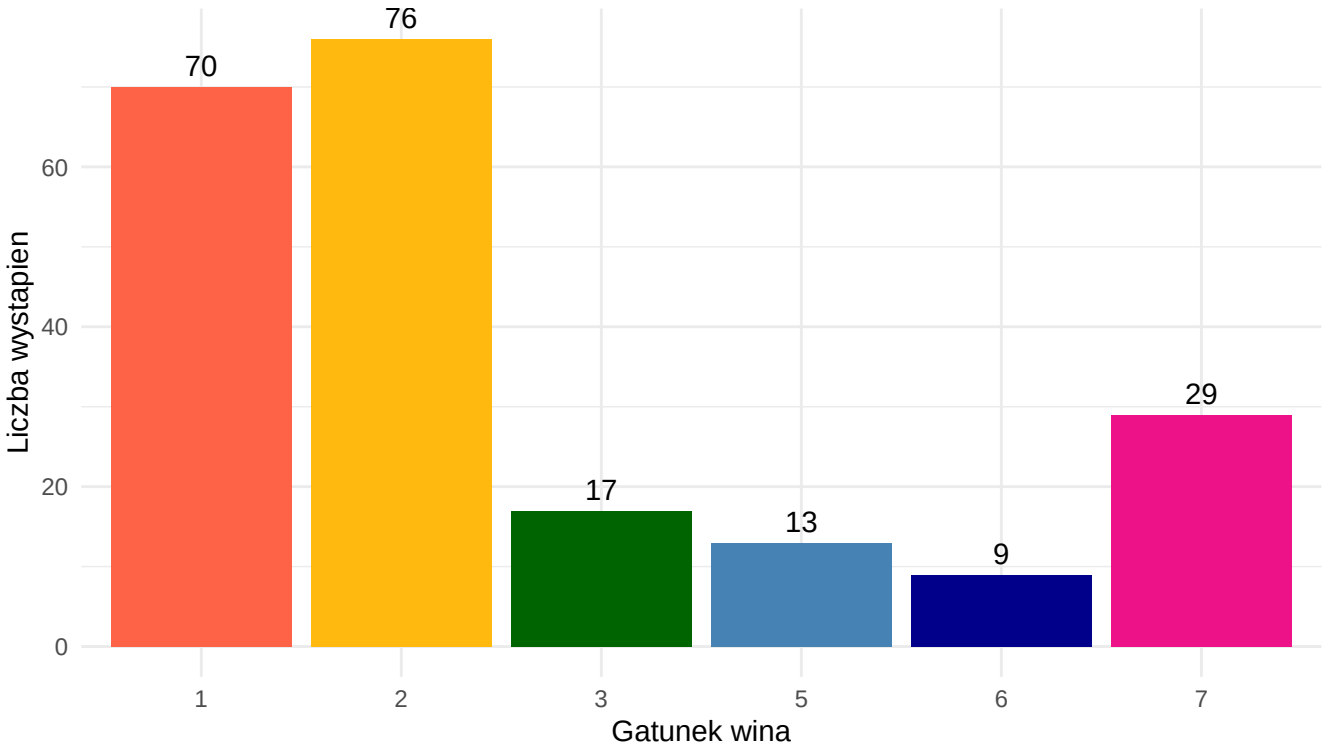
Zbiór danych glass zawiera **214** przypadków sześciu rodzajów szkła oraz **10** cech. Liczba brakujących danych wynosi **0**.

Znaczenie poszczególnych cech oraz ich typ przedstawiono w tabeli 23.

Tabela 23: Opis zmiennych w zbiorze danych wine

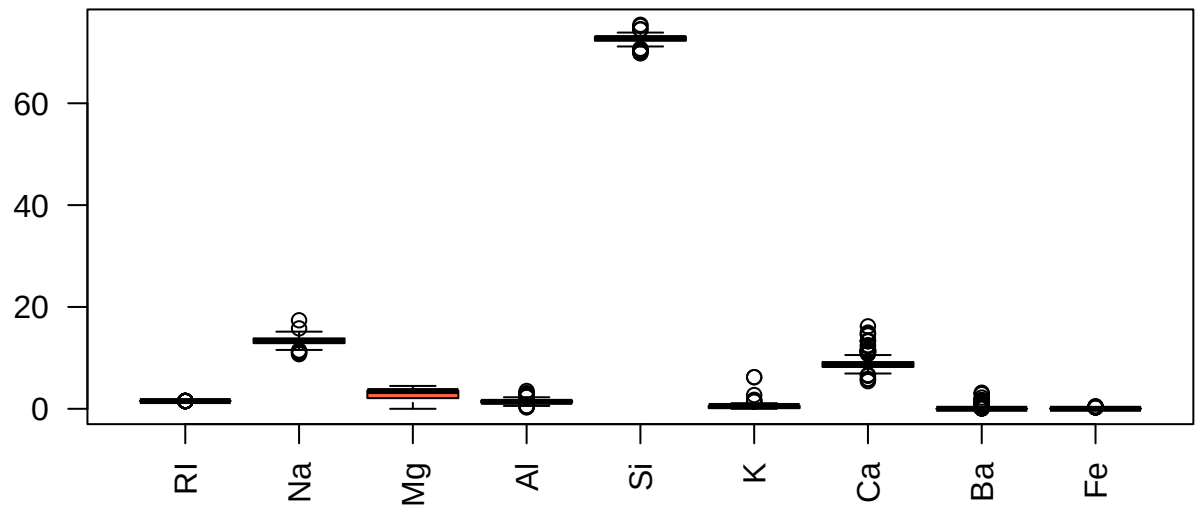
Zmienna	Typ	Opis
RI	numeric	Współczynnik załamania światła
Na	numeric	Zawartość azotu (procent wagowy w odpowiednim tlenku, podobnie jak atrybuty 3-9)
Mg	numeric	Zawartość magnezu
Al	numeric	Zawartość aluminium
Si	numeric	Zawartość krzemu
K	numeric	Zawartość potasu
Ca	numeric	Zawartość wapnia
Ba	numeric	Zawartość baru
Fe	numeric	Zawartość żelaza
Type	factor	Klasa (typ szkła: 1, 2, 3, 5, 6, 7)

W poszczególnych przypadkach sumy procentów wagowych znajdują się w zakresie 99.02-100.1. Nadmiar spowodowany jest najprawdopodobniej przez błędy przy zaokrąglaniu do dwóch miejsc po przecinku. Niedomiar natomiast może być spowodowany przez zawartość innych pierwiastków chemicznych, niezawartych w zbiorze.



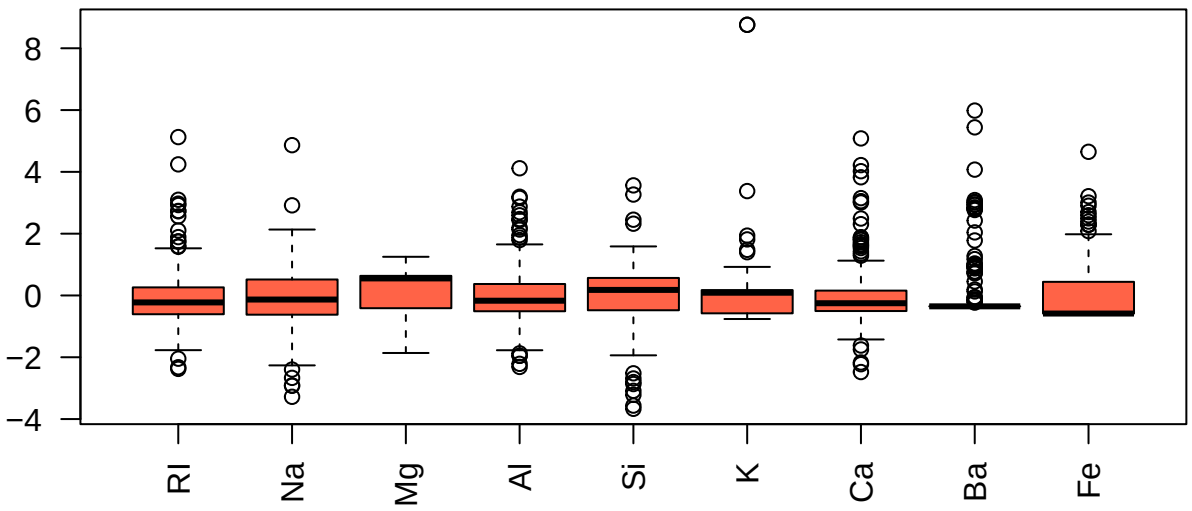
Wykres 2: Liczność poszczególnych typów szkła

Z Wykresu 1. można odczytać, że liczba obserwacji poszczególnych klas w zbiorze danych jest bardzo zróżnicowana. Siedemdziesiąt obserwacji lub więcej posiadają typy 1. oraz 2. (co stanowi 68.22% danych), reszta już po mniej niż 30 obserwacji.



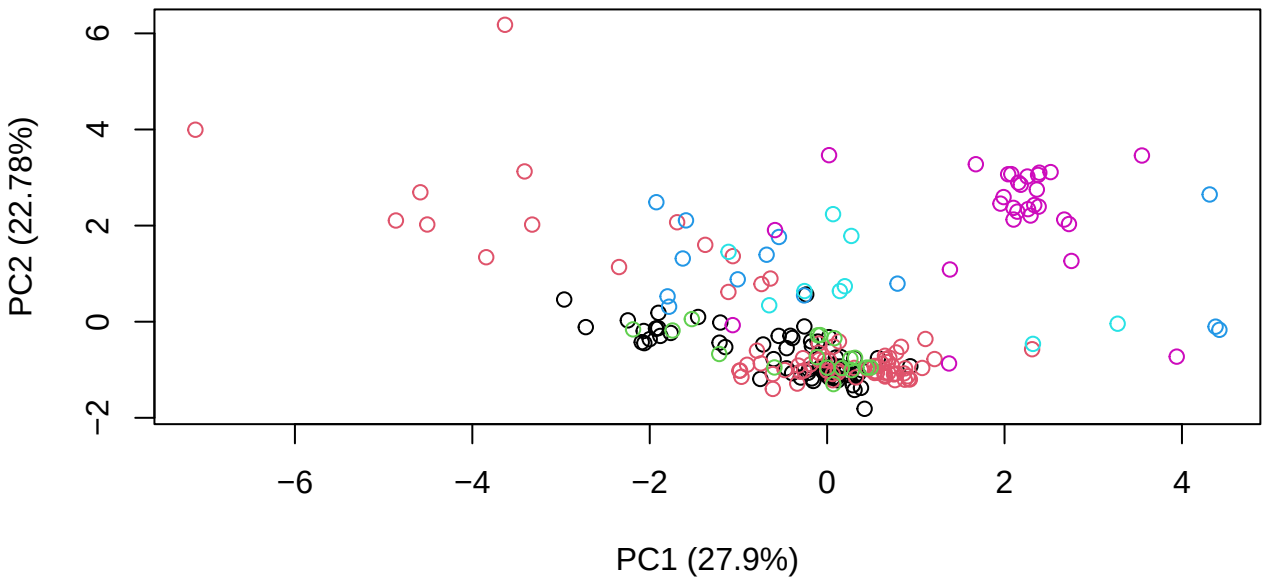
Wykres 3: Wykres pudełkowy, zmienne bez standaryzacji

KURWA STANDARYZACJA MACHEN



Wykres 4: Wykres pudełkowe, po standaryzacji

Teraz jest zajebiście



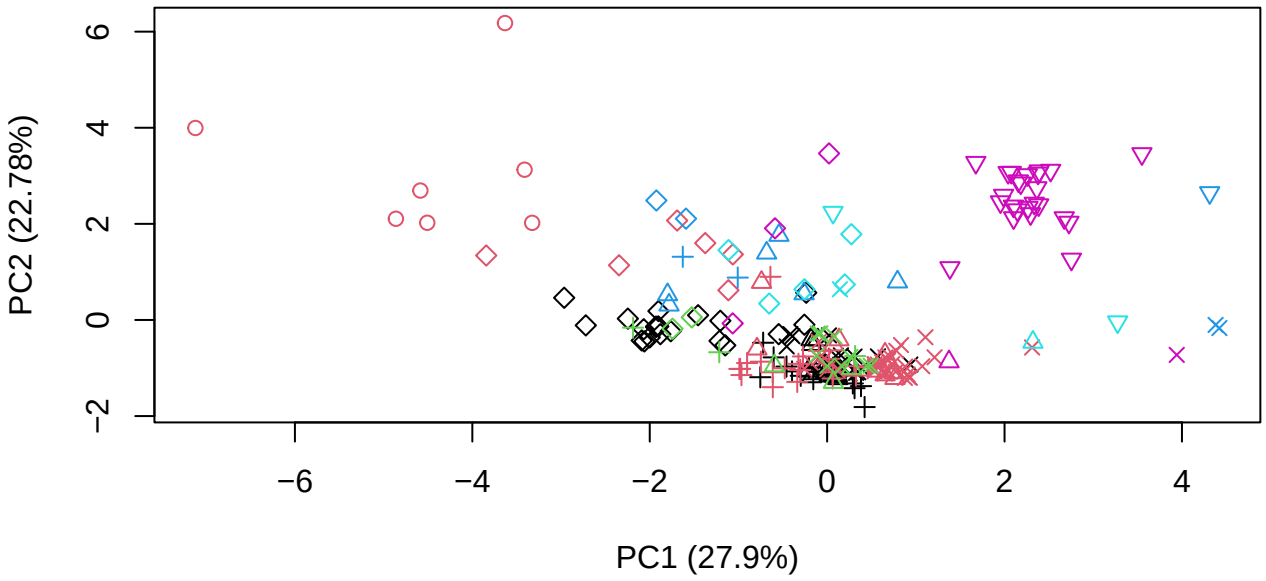
Wykres 5: Wizualizacja danych, PCA

Chuja widać, ciekawe kto wybrał ten zbiór?

2.2 Wyniki grupowania

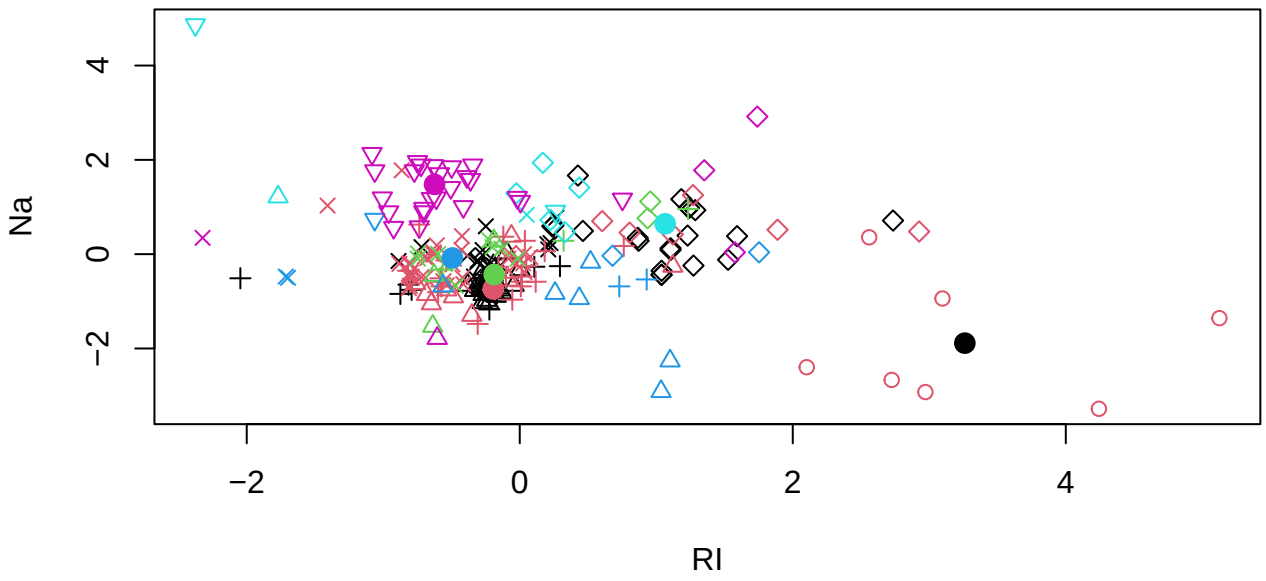
Przeprowadzamy dla rzeczywistej liczby etykiet, która wynosi **6**.

2.2.1 k-średnie



Wykres 6: PCA, kolory - rzeczywiste, kształt - wyniki

Gównianie mu poszło



Wykres 7: Wykres RI od Na, aby pokazać gdzie są wyznaczone centra skupień

Centra wywalone w kosmos, ale fajnie

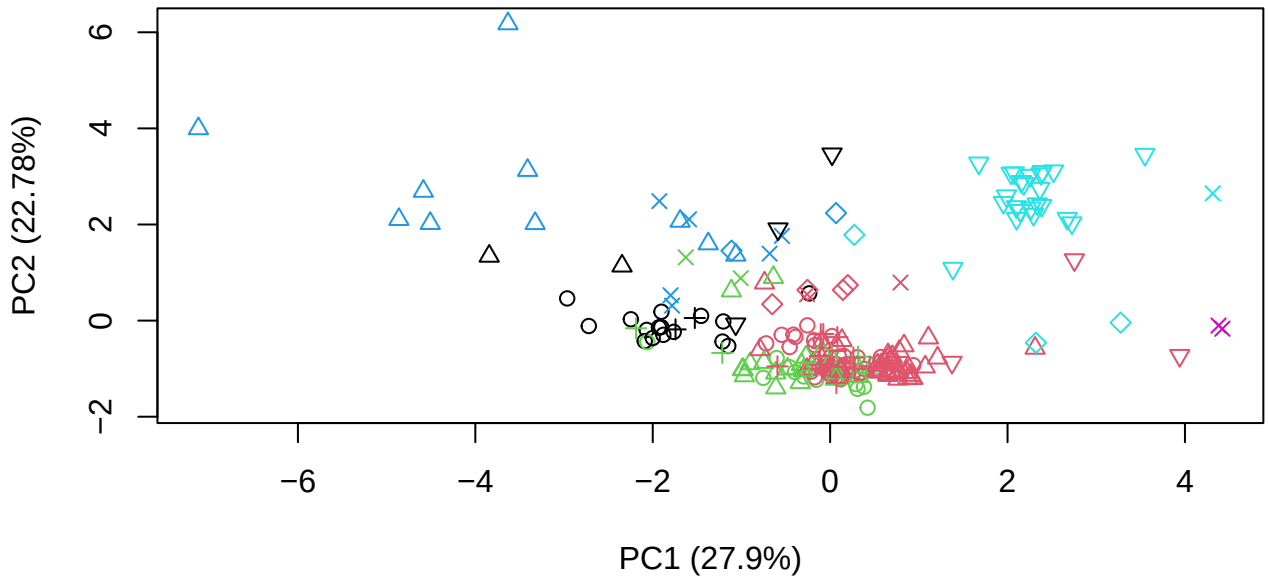
Tabela 24: Macierz błędów; metoda k-średnich

	1	2	3	5	6	7
1	20	6	2	2	5	3
2	17	34	9	2	1	1
3	13	20	3	2	0	0
5	20	9	3	6	1	1
6	0	7	0	0	0	0
7	0	0	0	1	2	24

Dokładność (macierz): 40.65%.

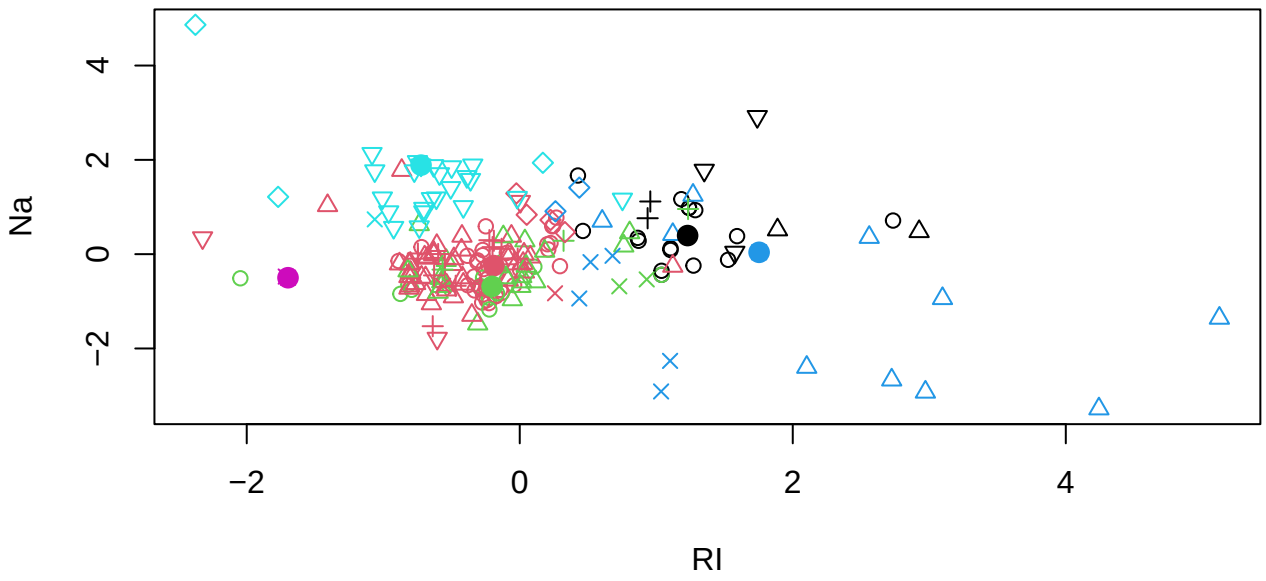
Dokładność (matchClasses, “exact”): 40.65%.

2.2.2 Partitioning Around Medoids (PAM)



Wykres 8: coś

Też słabo



Wykres 9: coś, z medoidami

Meh

Tabela 25: 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 26: 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

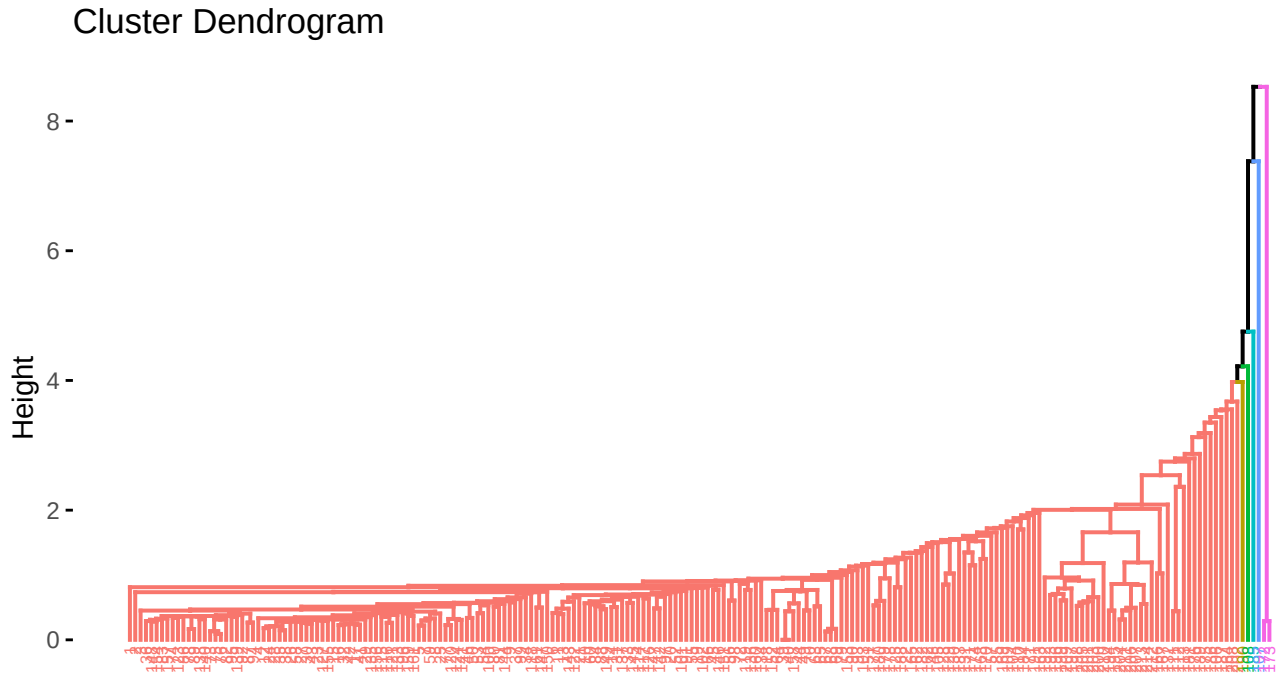
Dokładność: 33.64%.

Dokładność (matchClasses, “exact”): 42.99%.

Gorzej niż k-średnie *shocked emoji*.

2.2.3 Agglomerative Nesting (AGNES)

2.2.3.1 Najbliższy sąsiad



Wykres 10: AGNES: single linkage

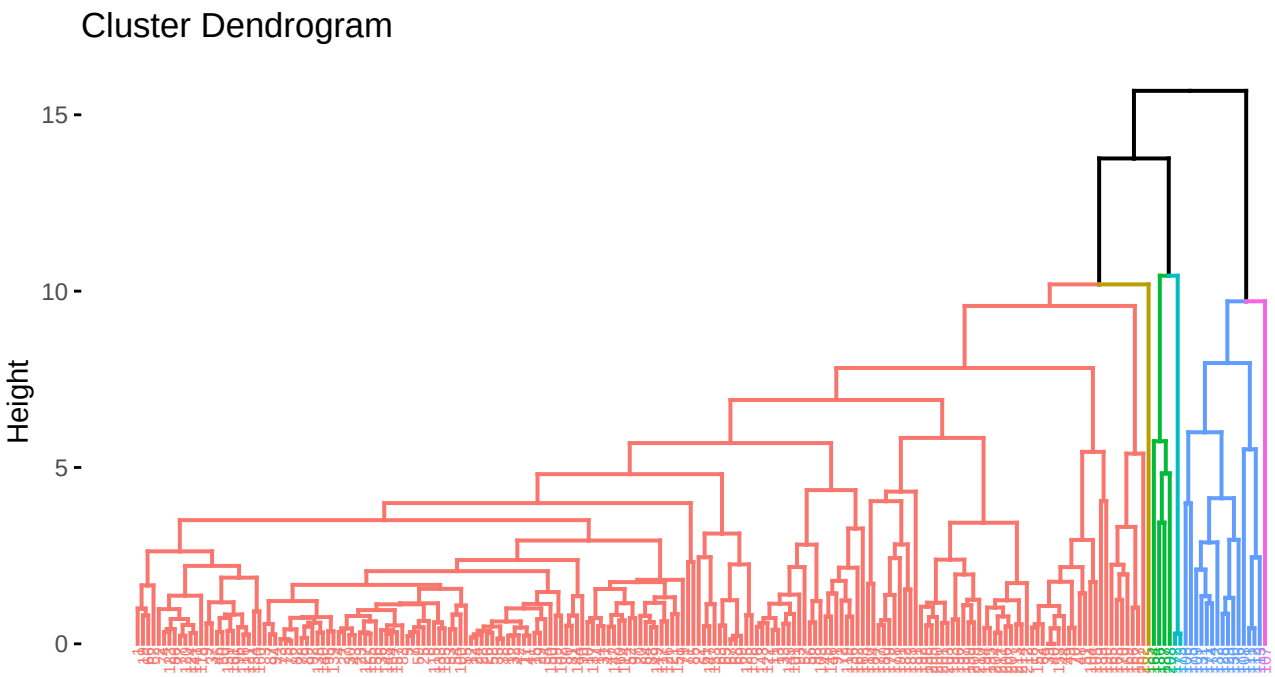
Tabela 27: Macierz błędów; agnes, najbliższy sąsiad

	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 11: AGNES: complete linkage

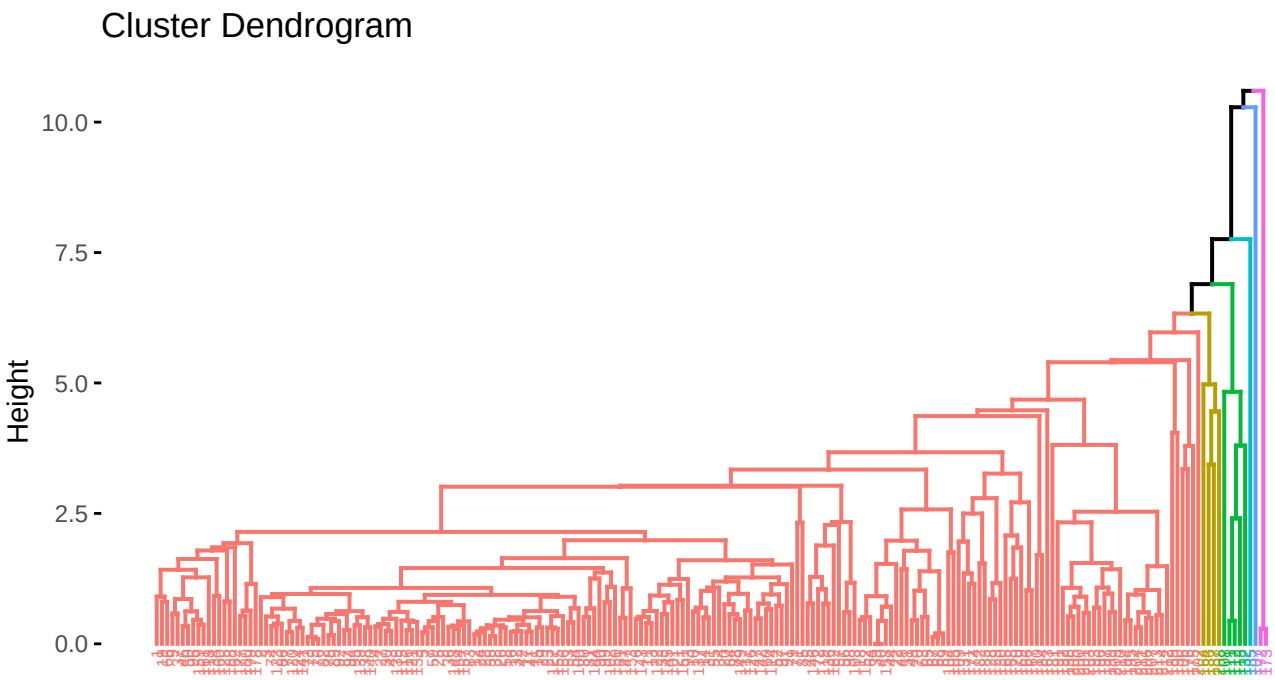
Tabela 28: Macierz błędów; agnes, najdalszy sąsiad

	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 12: AGNES: average linkage

Tabela 29: Macierz błędów; agnes, średnia odległość

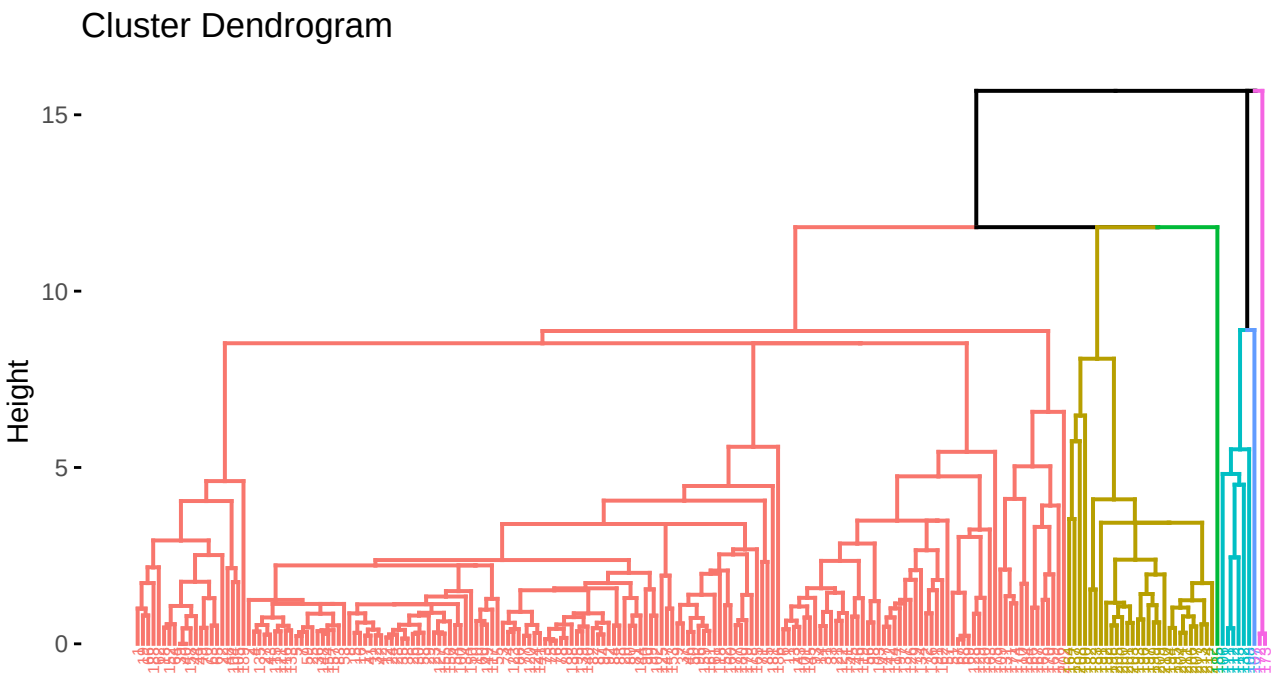
	1	2	3	5	6	7
1	70	70	17	10	8	26
2	0	1	0	0	0	0
3	0	5	0	0	0	0
5	0	0	0	1	0	3
6	0	0	0	2	0	0
7	0	0	0	0	1	0

Dokładność: 33.64%.

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

2.2.4 Divisive clustering (DIANA)

2.2.4.1 Odległość euklidesowa



Wykres 13: AGNES: average linkage

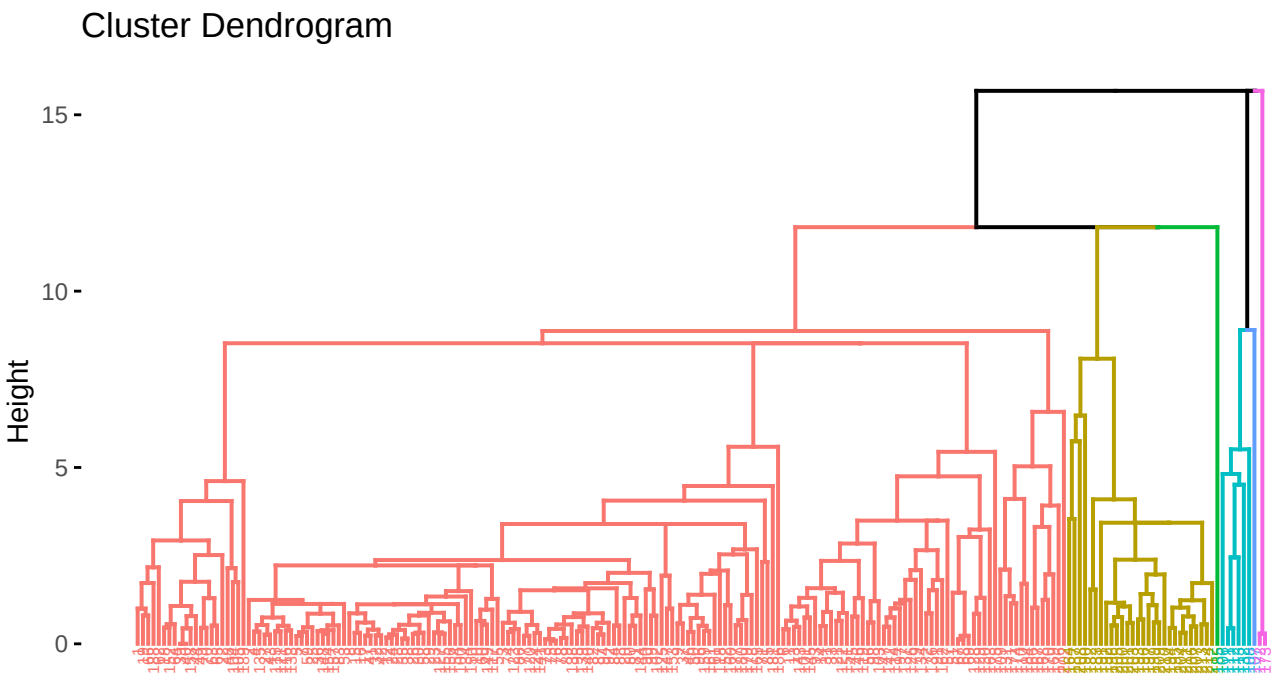
Tabela 30: Macierz błędów; agnes, średnia odległość

	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%.

2.2.4.2 Odległość Manhattan (taksówkowa)



Wykres 14: AGNES: average linkage

Tabela 31: 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%.

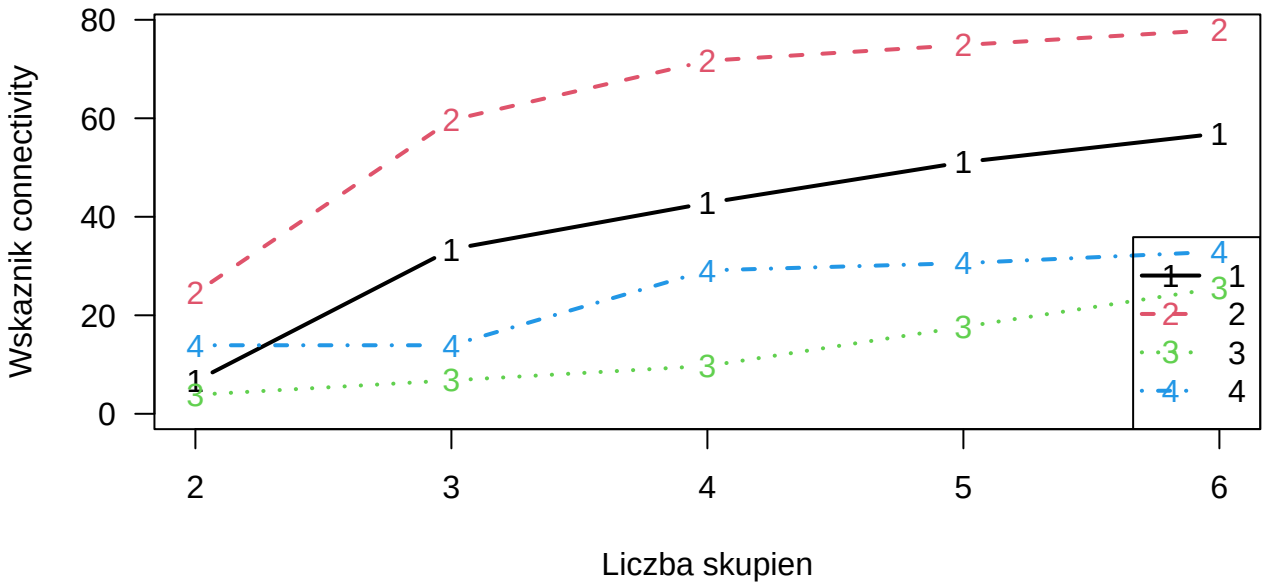
Dokładność (matchClasses, “exact”): 48.6%.

2.3 Ocena jakości grupowania i wizualizacja najlepszych wyników

2.3.1 Ocena

Legenda:

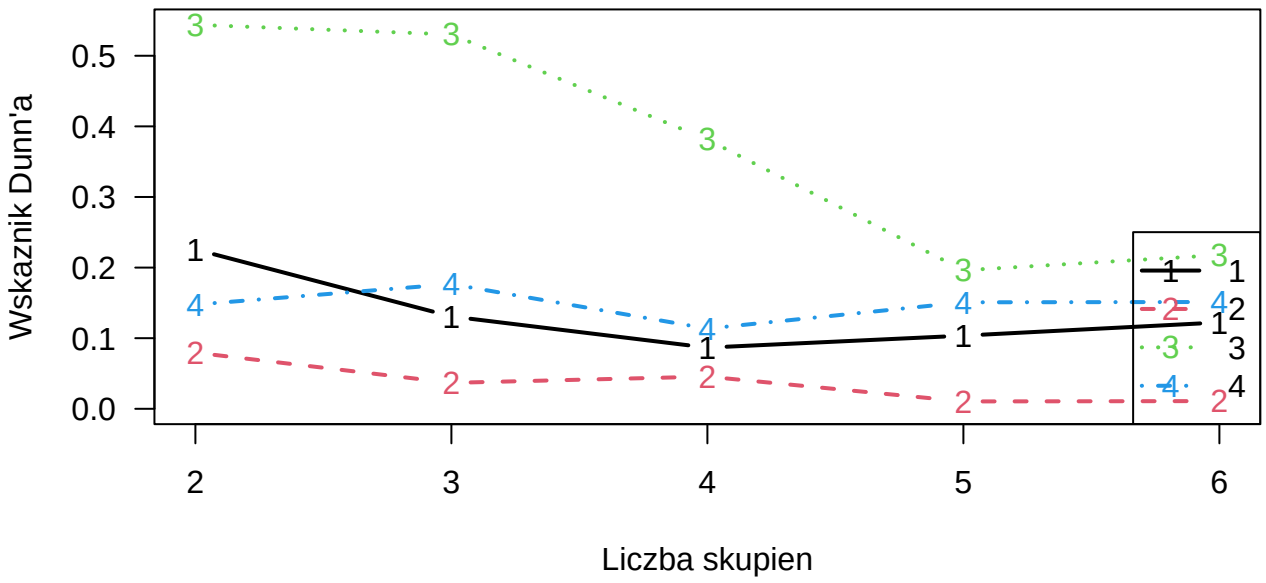
- 1 - kmeans,
- 2 - PAM,
- 3 - AGNES,
- 4 - DIANA.



Wykres 15: Connectivity

Connectivity - im mniejszy, tym lepszy:

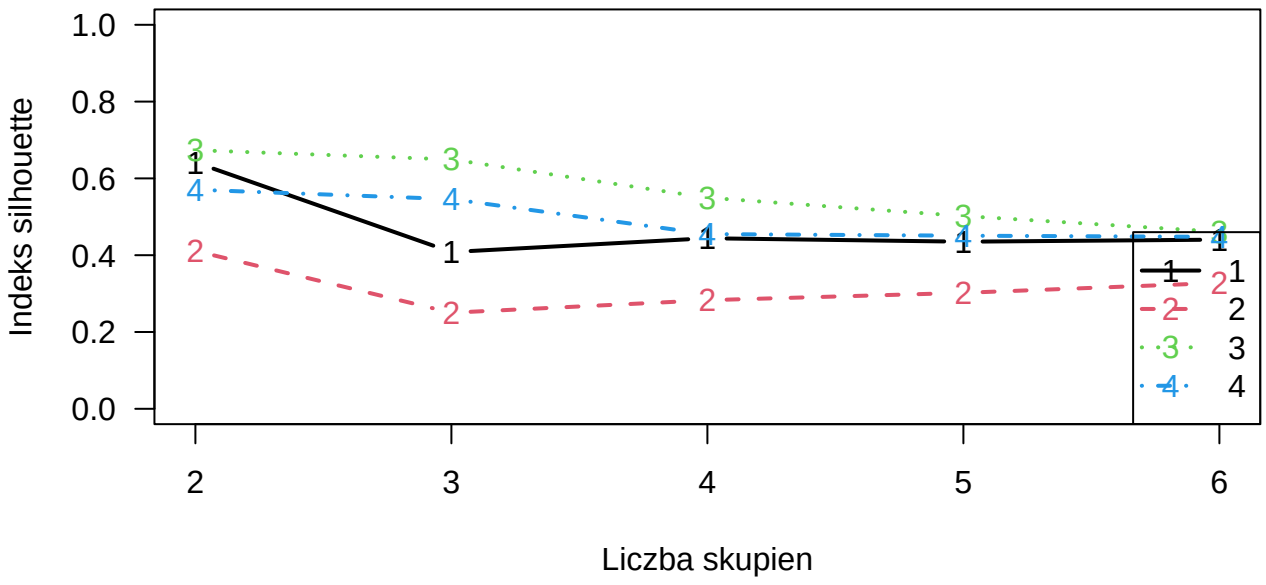
- kmeans - 2,
- PAM - 2,
- AGNES - 2 (najlepszy),
- DIANA - 3.



Wykres 16: Dunn

Dunn - im większy, tym lepszy:

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



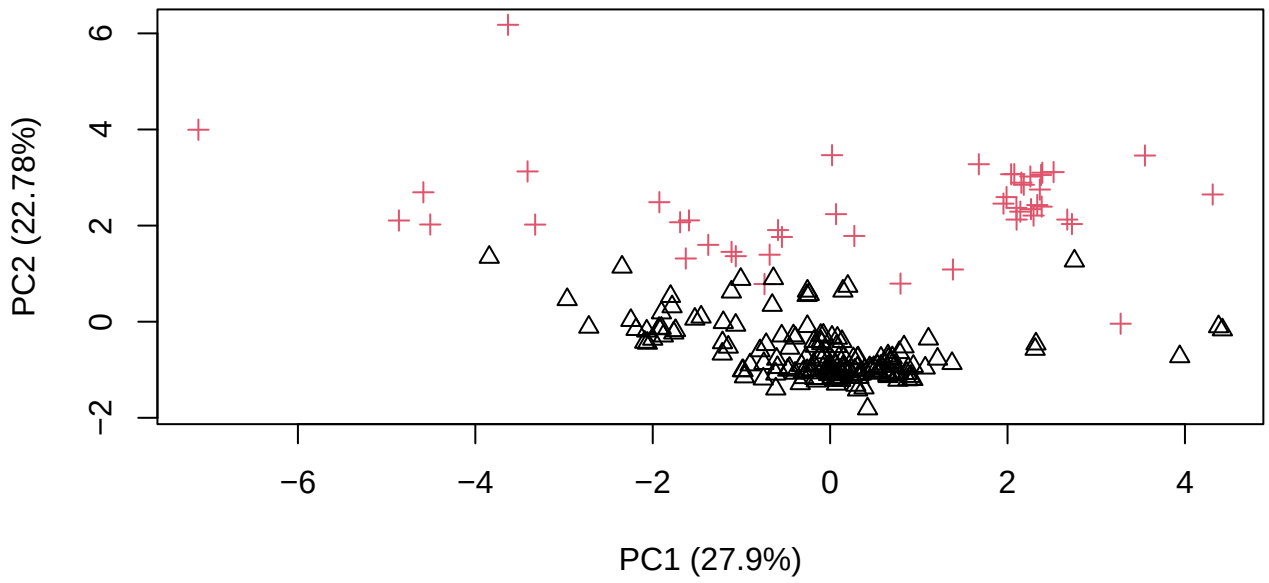
Wykres 17: Silhouette

Dunn - im większy, tym lepszy:

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

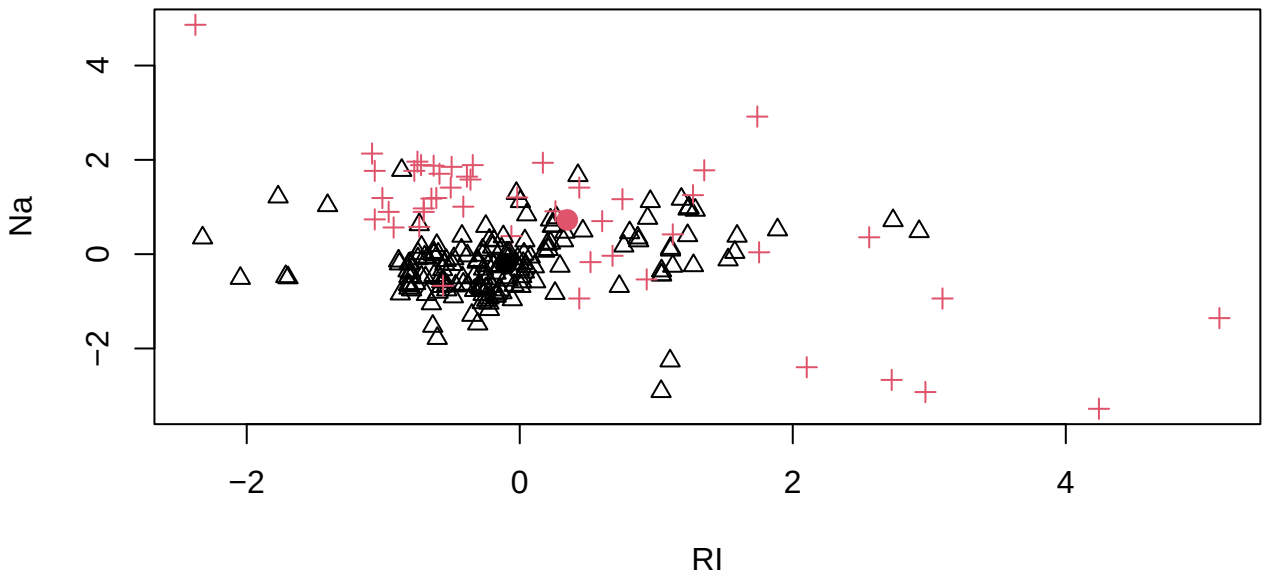
2.3.2 Wizualizacja

2.3.2.1 k-średnie



Wykres 18: PCA, kolory - rzeczywiste, kształt - wyniki, k=2

Gównianie mu poszło

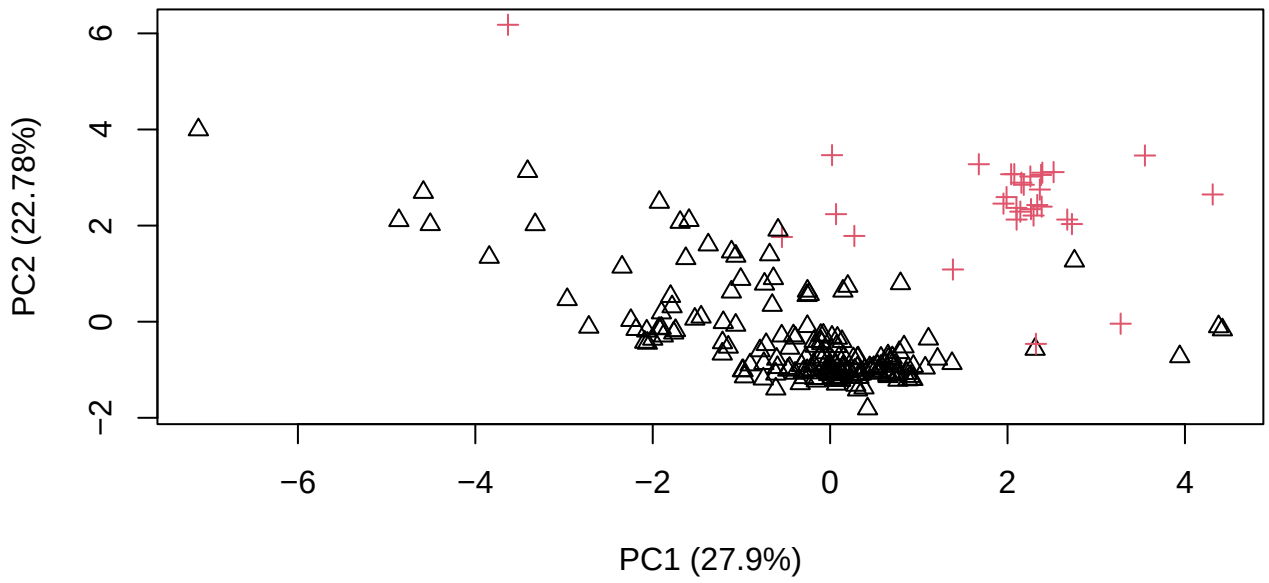


Wykres 19: Wykres RI od Na, aby pokazać gdzie są wyznaczone centra skupień, k=2

Centra wywalone w kosmos, ale fajnie

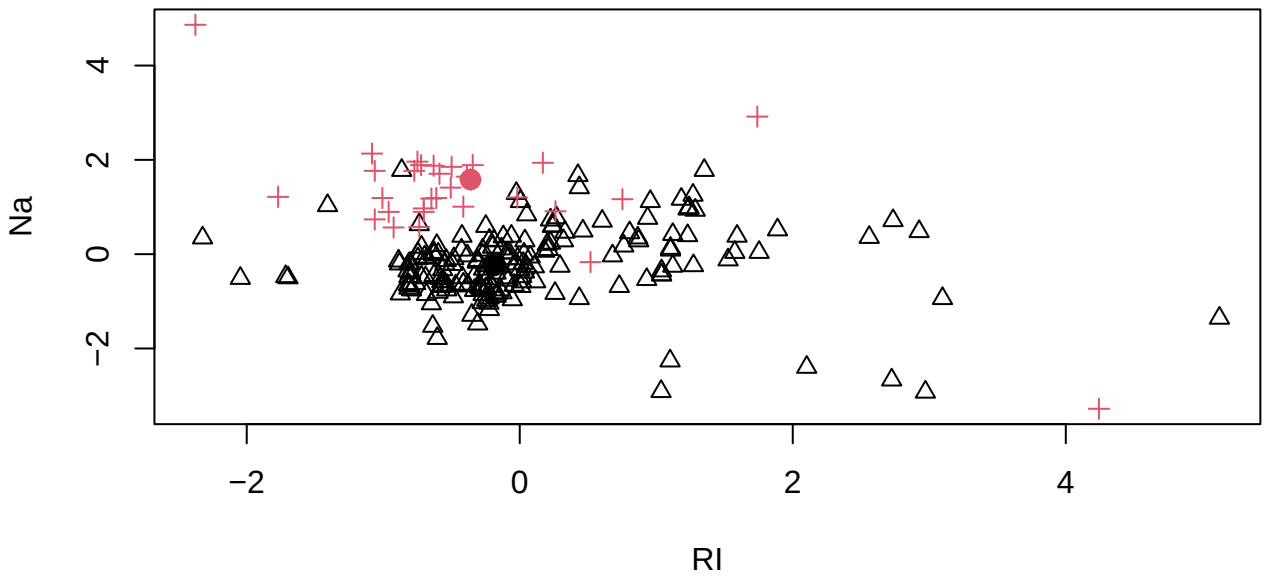
Dokładność (matchedClasses): 95.96%.

2.3.2.2 PAM



Wykres 20: coś, k=2

Też słabo



Wykres 21: coś, z medoidami, k=2

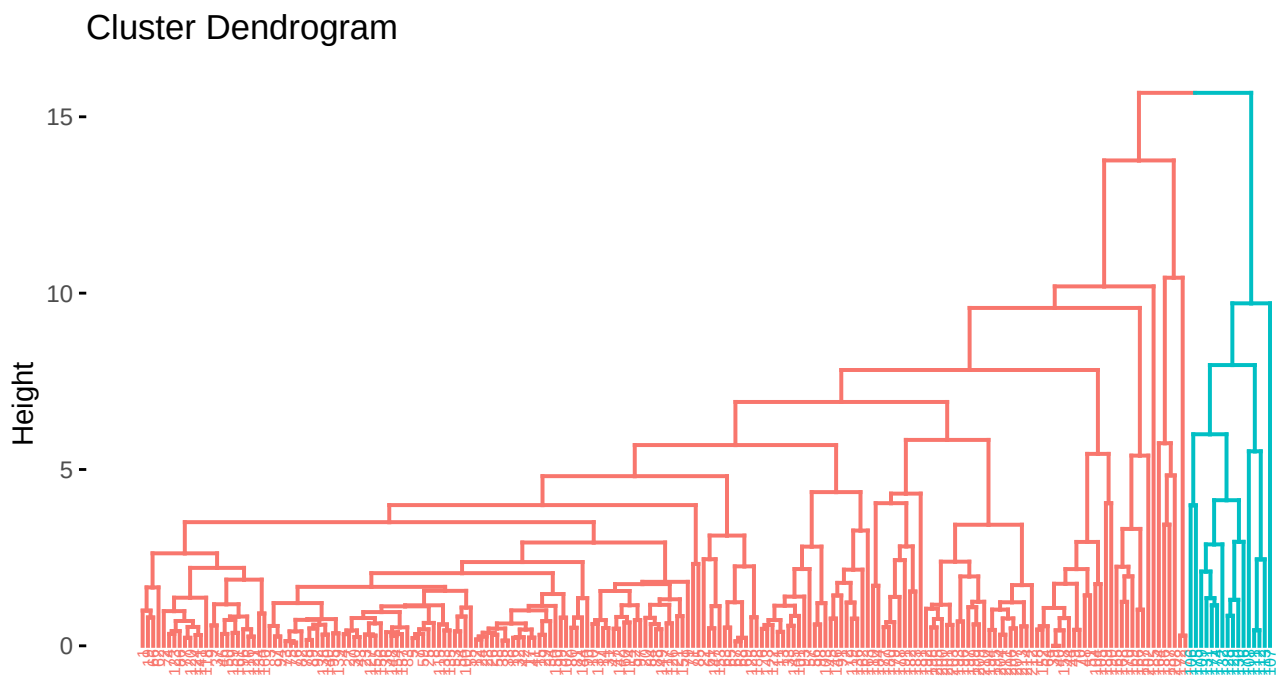
Meh

Dokładność (matchedClasses): 94.29%.

Tabela 32: 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

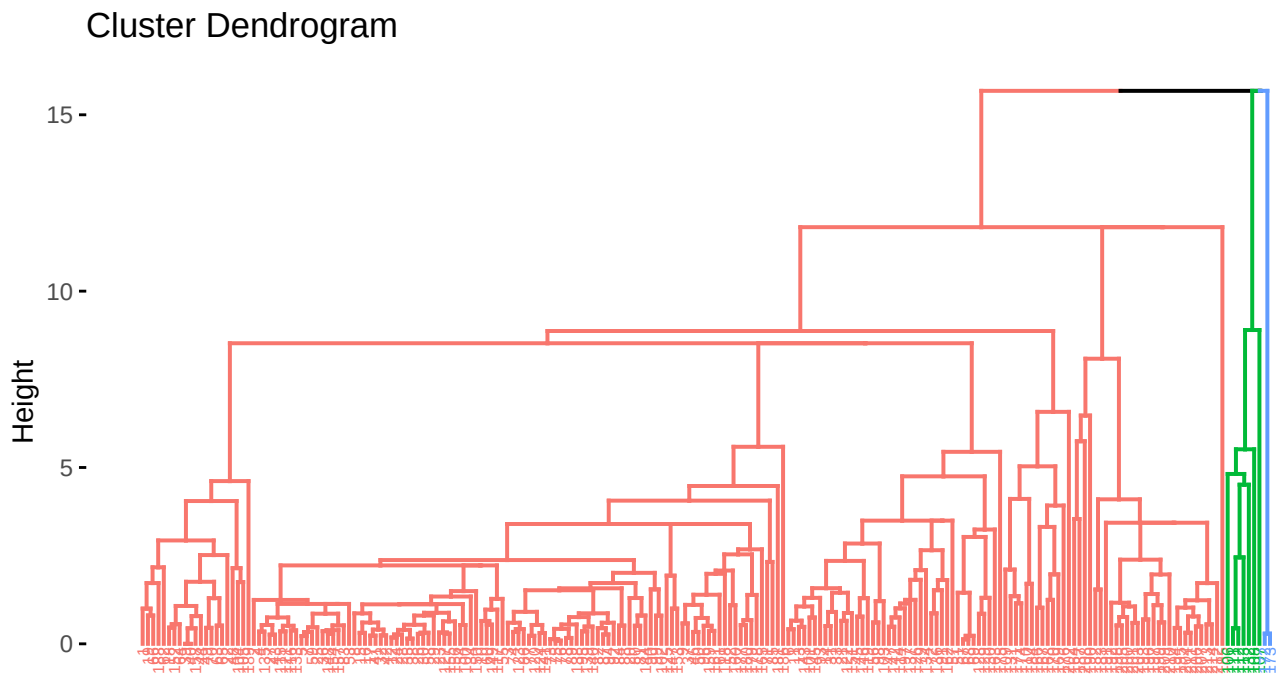


Wykres 22: AGNES: complete linage, k=2

Tylko dla najdalszych sąsiadów, bo dał najlepsze wyniki.

Dokładność (matchedClasses): 87.64%.

2.3.2.4 DIANA



Wykres 23: DIANA, k=3

Dokładność (matchedClasses): 49.69%.

2.4 Wnioski

3 Podsumowanie

PS. Czas wykonywania kodu wynosi 10 minut i 28 sekund.