Klasyfikacja na bazie modelu regresji liniowej oraz porównanie metod klasyfikacji na podstawie danych iris, Glass

Eksploracja danych

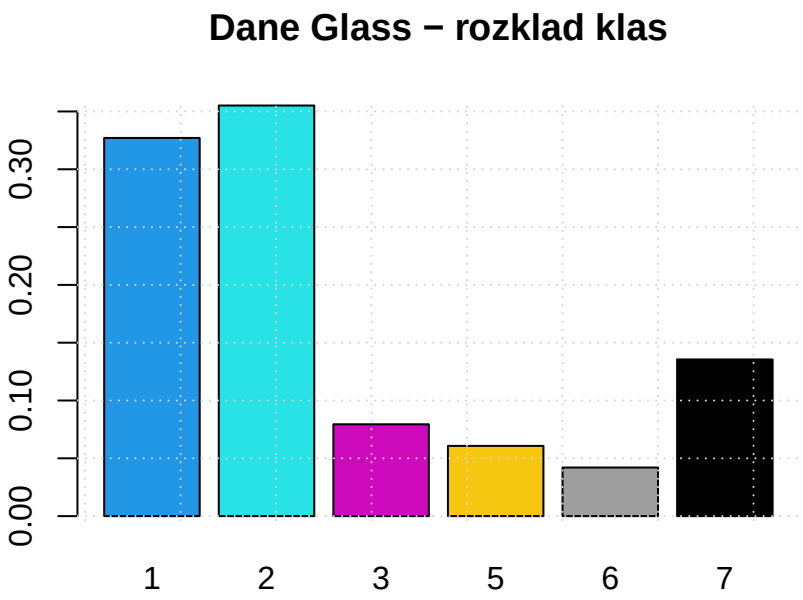
Tomasz Warzecha, album 282261

2025-06-19

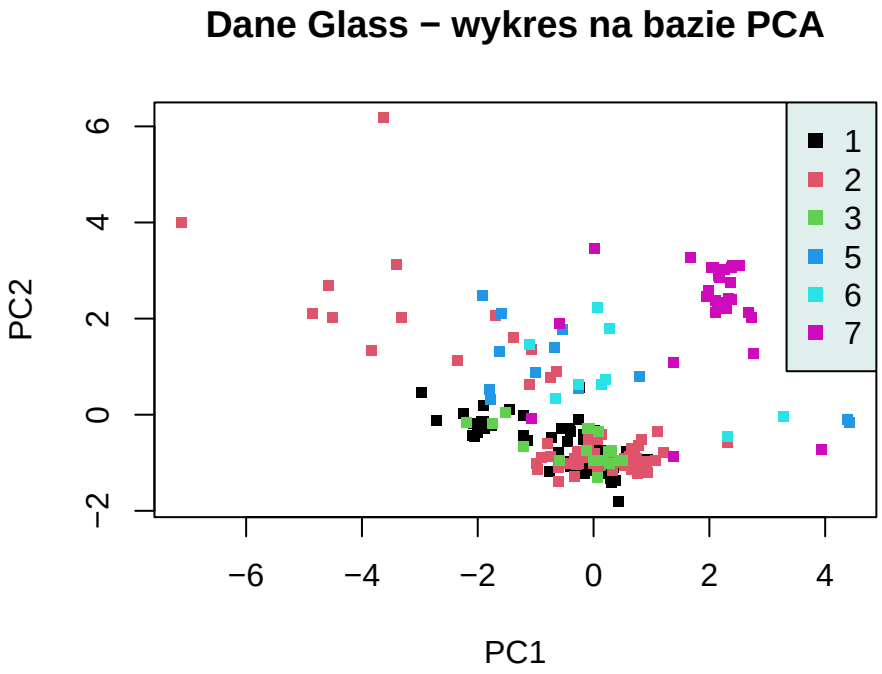
Spis treści

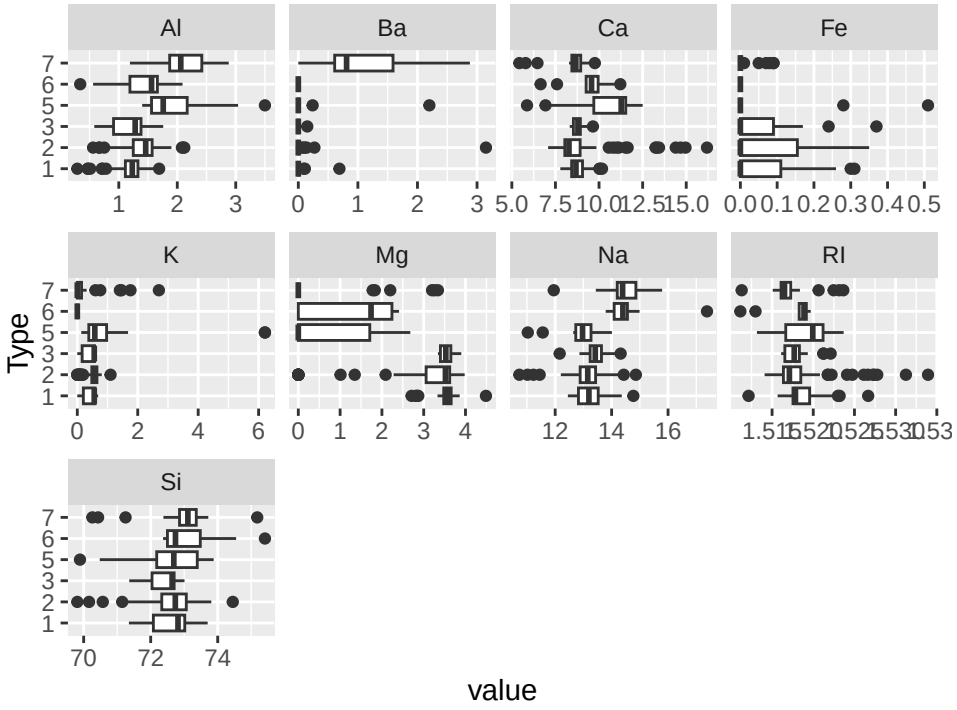
1 Zaawansowane metody klasyfikacji

1.1 Rodziny klasyfikatorów/uczenie zespołowe

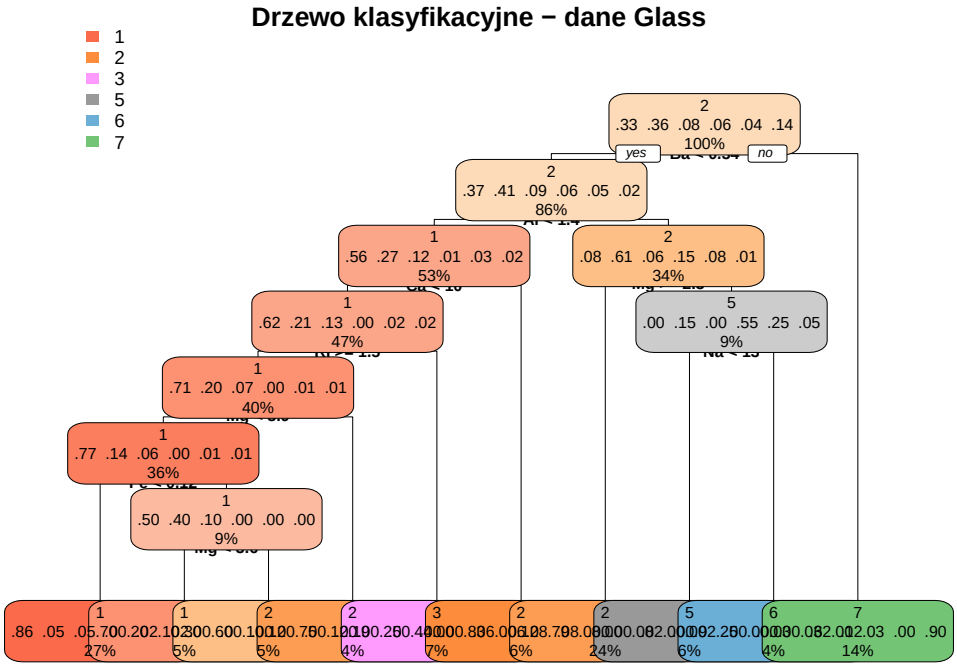


Importance of components: PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 Standard deviation 1.585 1.4318 1.1853 1.0760 0.9560 0.72639 0.6074 0.25269 Proportion of Variance 0.279 0.2278 0.1561 0.1286 0.1016 0.05863 0.0410 0.00709 Cumulative Proportion 0.279 0.5068 0.6629 0.7915 0.8931 0.95173 0.9927 0.99982 PC9 Standard deviation 0.04011 Proportion of Variance 0.00018 Cumulative Proportion 1.00000


Dane Glass − wykres na bazie PCA
PC2
PC1
1
2
3
5
6
7


Al
Ba
Ca
Fe
K
Mg
Na
RI
Si
Type
value

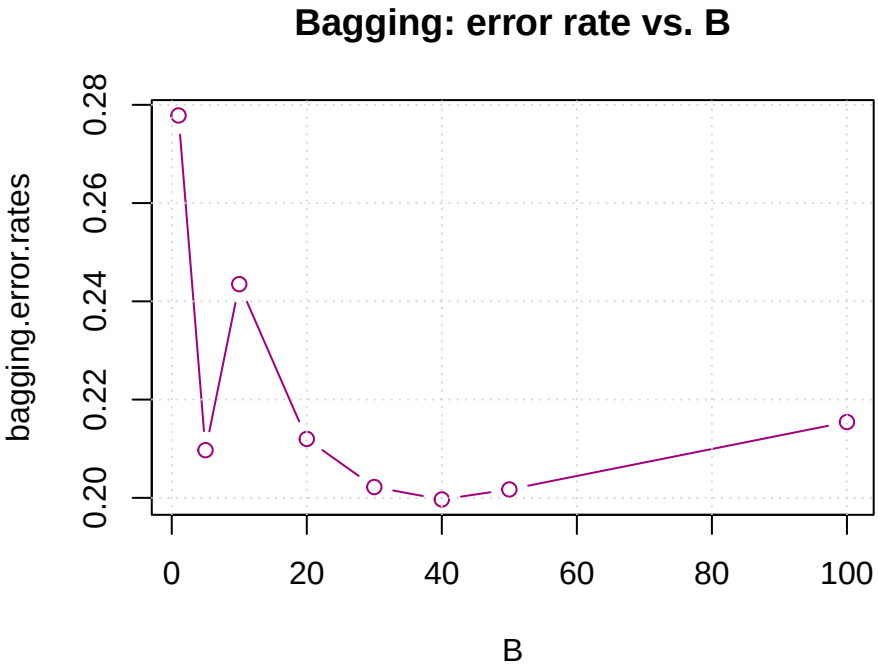
1.1.1 Pojedyńcze drzewo klasyfikacyjne



Na powyższym Rysunku przedstawione zostało drzewo klasyfikacyjne dla całego zbioru danych. Poniżej znajdują się dane ile wynosiły błędy klasyfikacyjne dla zbioru testowego:

Błąd klasyfikacji - zbiór uczący: 0.204 Błąd klasyfikacji - zbiór testowy: 0.361

1.1.2 Metoda bagging



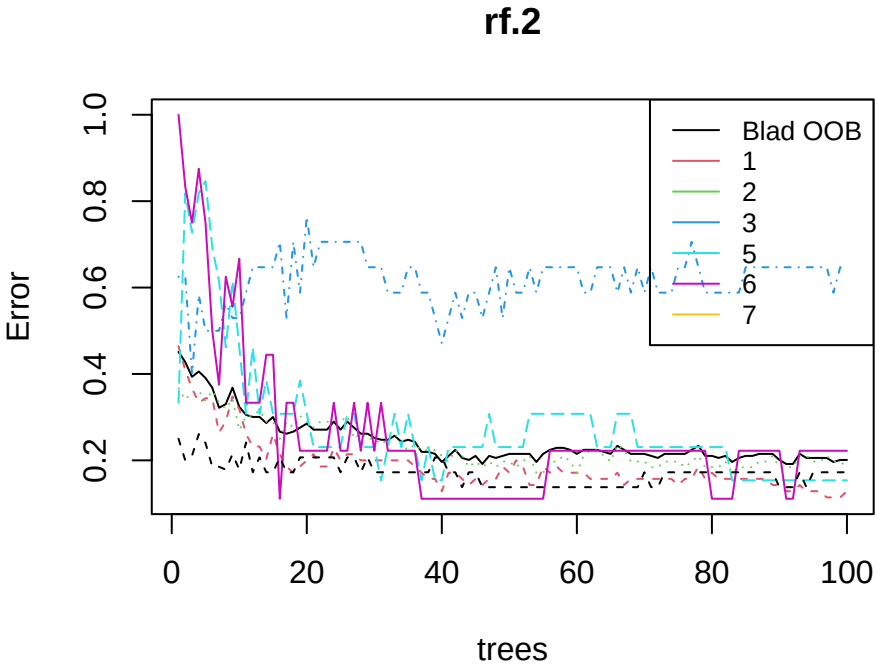
Można zauważyć na Rysunku, że wraz ze wzrostem liczby drzew (parametru B) błąd kla-

syfikacji początkowo gwałtownie maleje – już przy 5–10 drzewach osiąga znaczną poprawę. Najniższy poziom błędu przypada na około 40 drzew. Dalsze zwiększanie liczby drzew nie przynosi wyraźnej poprawy, a wręcz może prowadzić do lekkiego pogorszenia wyników, jak w przypadku B=100. Wskazuje to, że optymalna liczba drzew mieści się w przedziale około 30–50

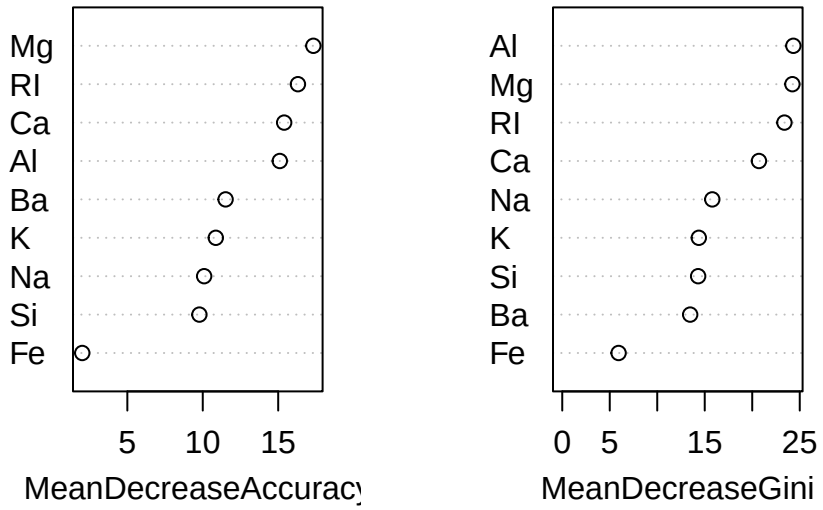
1.1.3 Metoda Random Forest

```
real.labels
```

pred.labels 1 2 3 5 6 7 1 70 0 0 0 0 0 2 0 76 0 0 0 0 3 0 0 17 0 0 0 5 0 0 0 13 0 0 6 0 0 0 0 9 0 7 0 0 0 0 0 29 1 2 3 5 6 7 class.error 1 61 6 2 0 0 1 0.1285714 2 10 62 1 2 1 0 0.1842105 3 9 2 6 0 0 0 0.6470588 5 0 1 0 11 0 1 0.1538462 6 1 1 0 0 7 0 0.2222222 7 2 3 0 0 0 24 0.1724138



Wykres błędu dla metody Random Forest wskazuje, że model osiąga stabilność przy około 40 drzewach — dalsze zwiększanie ich liczby nie prowadzi już do istotnej poprawy wyników. Model lepiej radzi sobie z klasyfikacją przypadków niechorobowych, natomiast wykrywanie przypadków choroby wypada gorzej. Taka asymetria może być skutkiem nierównomiernego rozkładu klas w zbiorze treningowym. Warto rozważyć zastosowanie technik wyrównywania klas, takich jak oversampling czy undersampling, lub alternatywnych miar oceny skuteczności modelu, które lepiej uwzględniają niezrównoważone dane. Szczególnie trudna okazuje się klasyfikacja klasy 5, dla której błąd nie spada poniżej 50%.



Na wykresie przedstawiono istotność zmiennych w modelu Random Forest. Można zauważyć, które cechy mają największy wpływ na przyporządkowanie obserwacji do odpowiednich klas. W przypadku miary spadku dokładności (MeanDecreaseAccuracy) kluczową rolę odgrywa zmienna Mg, a nieco mniejsze znaczenie ma zmienna RI. Z kolei według miary spadku nieczystości Gini (MeanDecreaseGini), największy wkład w budowę modelu mają zmienne Al oraz Mg, które wykazują podobny poziom istotności.

1.1.4 Wnioski

W przypadku zastosowania pojedynczego drzewa klasyfikacyjnego błąd klasyfikacji na zbiorze testowym wyniósł około 36%, co wskazuje na stosunkowo niską skuteczność tego podejścia dla danych typu Glass.

Zastosowanie metody baggingu znacząco poprawiło jakość klasyfikacji – już przy około 5–10 drzewach odnotowano zauważalny spadek błędu, który osiągnął minimum w okolicach 40 drzew. Dalsze zwiększanie liczby replikacji nie przynosiło już wyraźnych korzyści, a w niektórych przypadkach powodowało nawet lekkie pogorszenie wyników. Optymalna liczba drzew mieściła się w przedziale 30–50.

Metoda Random Forest również pozwoliła na poprawę jakości klasyfikacji. Wykres błędu OOB wskazał, że model stabilizuje się przy około 40 drzewach. Jednakże, analiza macierzy błędów pokazała wyraźną asymetrię skuteczności modelu — klasy były rozpoznawane z różną dokładnością. Szczególnie trudna okazała się klasyfikacja klasy 5, dla której błąd nie spadał poniżej 50%. Może to wynikać z nierównomiernego rozkładu klas w zbiorze treningowym.

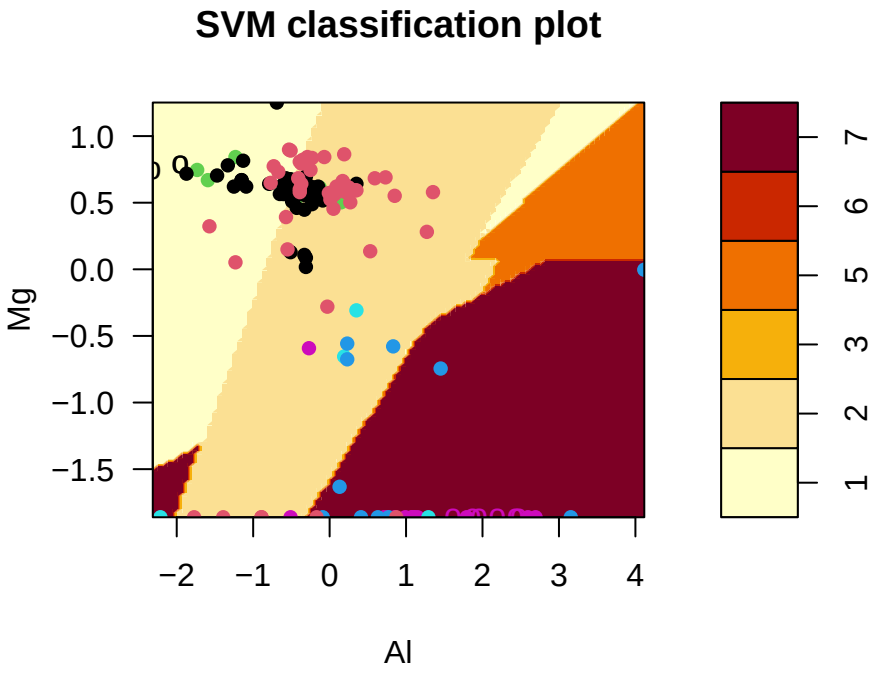
Podsumowując, najlepsze rezultaty uzyskano za pomocą metody bagging, która pozwoliła na największą redukcję błędu klasyfikacyjnego względem pojedynczego drzewa. Choć Random Forest również poprawił skuteczność klasyfikacji, to jego działanie było mniej efektywne przy

obecnym rozkładzie danych.

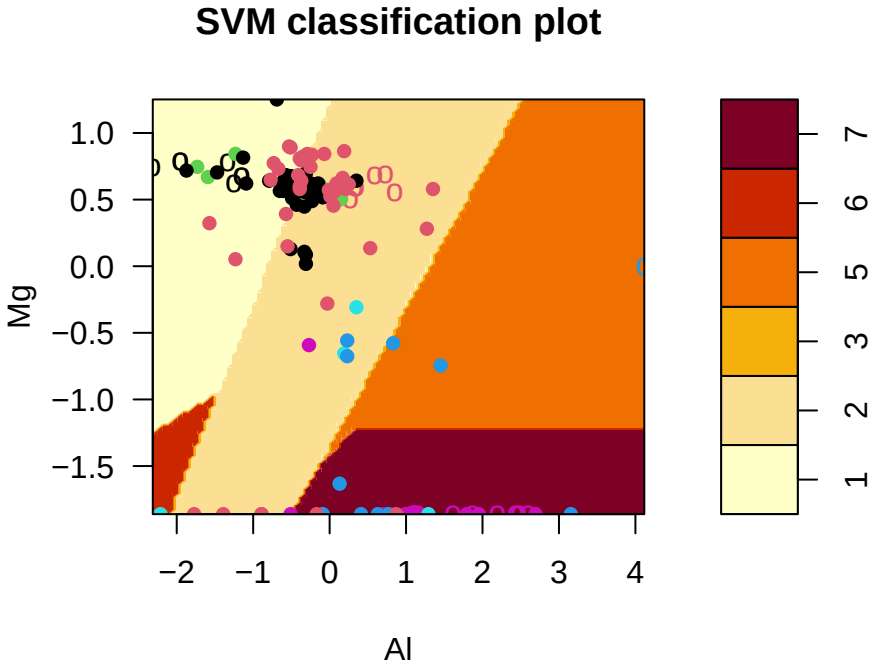
1.2 Metoda wektorów nośnych (SVM)

Metodę wektorów nośnych zastosowano dla dwóch zmiennych o najlepszych zdolności dyskryminacyjnych - Mg oraz Al, aby dobrze zwizualizować je na wykresach.

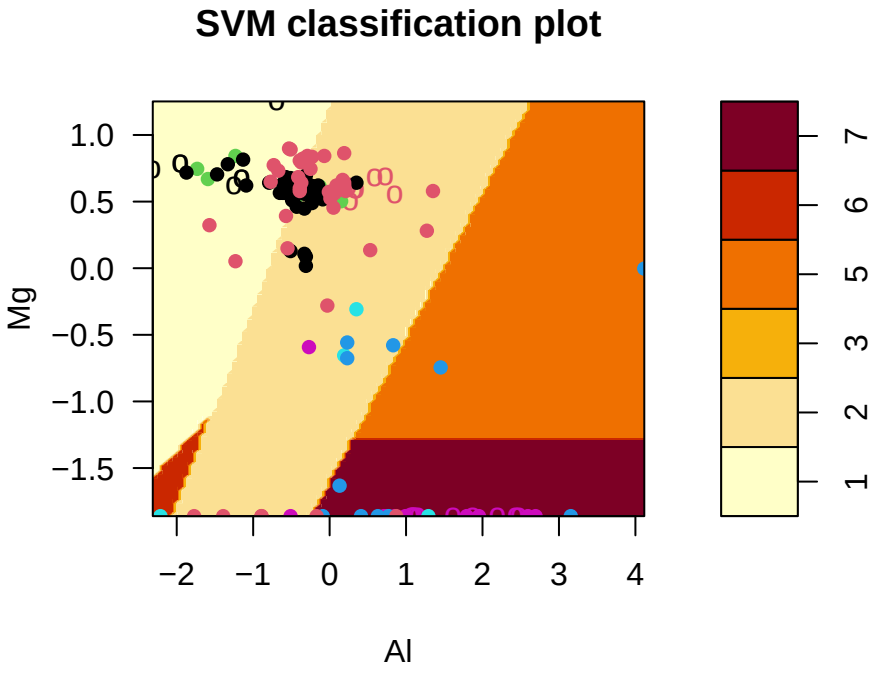
1.2.1 Porównanie skuteczności funkcji jądrowych i parametru kosztów



Skuteczność klasyfikacji: 0.486

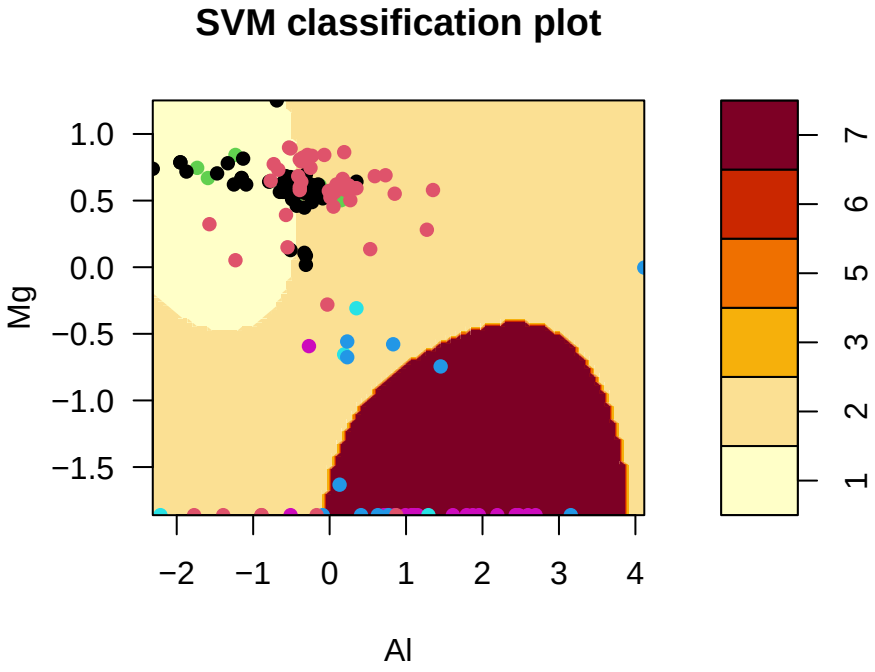


Skuteczność klasyfikacji: 0.514

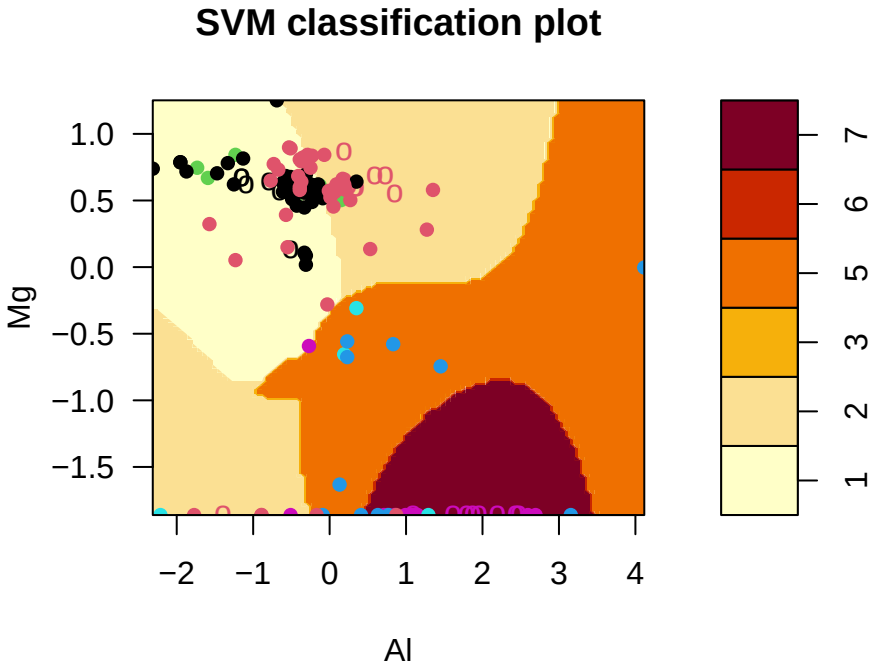


Skuteczność klasyfikacji: 0.514

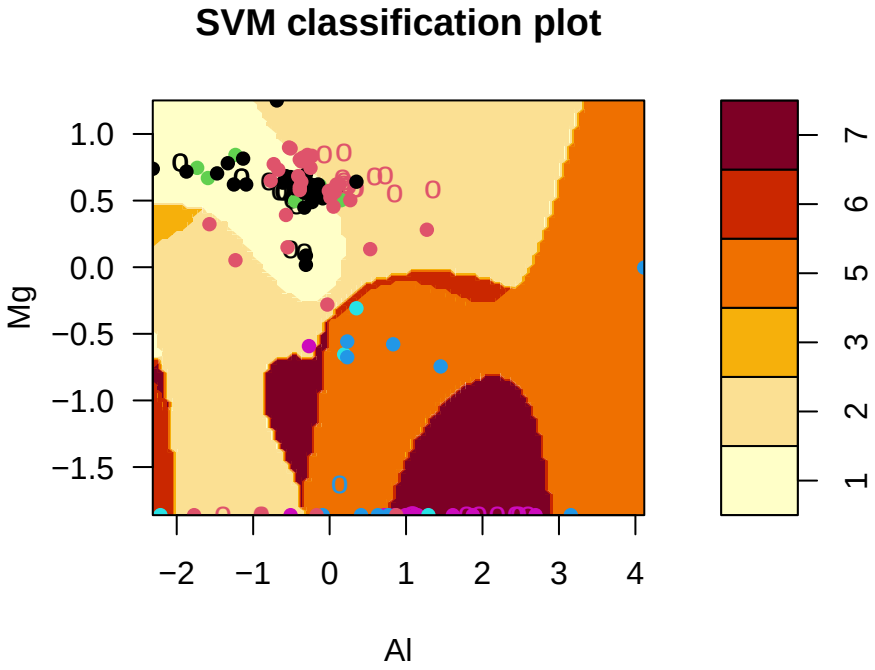
Na wykresach powyżej przedstawiono funkcje decyzyjne metody SVM wyznaczone na podstawie dwóch zmiennych: Mg oraz Al. Analiza skuteczności klasyfikacji dla różnych wartości parametru regularyzacji C pokazuje, że lepsze wyniki osiągnięto przy wyższych wartościach tego parametru. Zarówno dla C=1, jak i C=10 skuteczność klasyfikacji wyniosła 51,4%, natomiast dla niższego C=0,1 była nieco niższa i wyniosła 48,6%. Choć różnice nie są duże, wskazuje to, że zwiększenie wartości C poprawia zdolność modelu do dopasowania się do danych, kosztem mniejszego marginesu.



Skuteczność klasyfikacji: 0.514



Skuteczność klasyfikacji: 0.528



Skuteczność klasyfikacji: 0.542

Na wykresach powyżej przedstawiono funkcje decyzyjne SVM dla zmiennych Mg oraz Al, wykorzystując jądro radialne oraz różne wartości parametru kosztu C. Najlepszą skuteczność uzyskano dla parametru C równego 10, wynoszącą 54,2%. Niższe wyniki odnotowano dla C = 1 (52,8%) oraz najniższy dla C = 0,1 (51,4%).

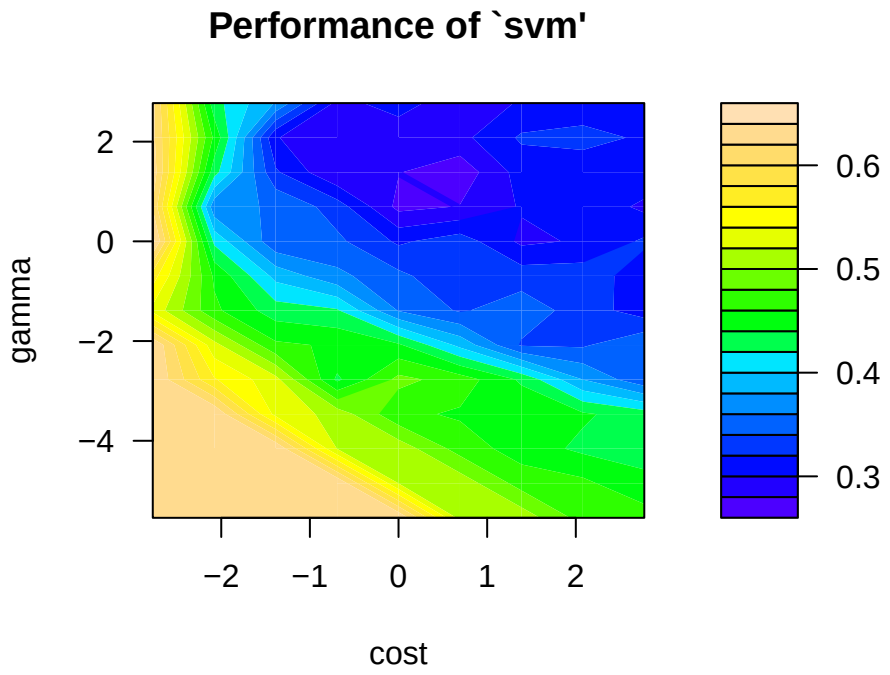
Skuteczność klasyfikacji zależy zarówno od doboru funkcji jądrowej, jak i wartości parametru kosztu C. Choć różnice w wynikach nie są znaczne, odpowiednia konfiguracja parametru C może prowadzić do zbliżonych rezultatów. Mimo to, przy identycznych ustawieniach, lepsze efekty uzyskano przy zastosowaniu jądra radialnego.

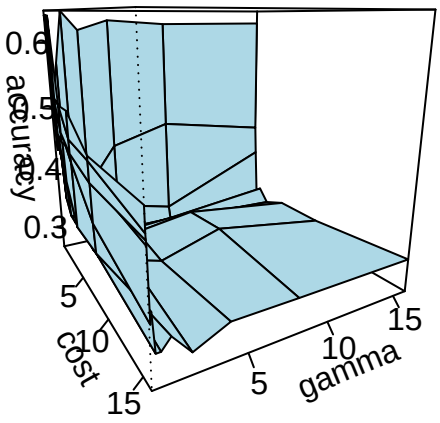
1.2.2 Dobranie najlepszych parametrów dla jądra radialnego

Poniższe wykresy przedstawiają dokładność klasyfikacji w zależności od parametru gamma oraz cost (parametr oznaczony jako C).

Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation
- best parameters: cost gamma 2 4
- best performance: 0.2671429





Skuteczność klasyfikacji: 0.556

Najlepszymi parametrami, uzyskanymi dzięki odpowiedniej optymalizacji, okazały się koszt C = 2 oraz gamma = 4. Skuteczność klasyfikacji przy tych ustawieniach wyniosła 55,6%. Jest to najlepszy wynik osiągnięty przy zastosowaniu funkcji jądrowych (jądro radialne) w metodzie wektorów nośnych. Można zatem stwierdzić, że właściwa optymalizacja pozwoliła poprawić skuteczność klasyfikacji.

1.3 Wnioski końcowe

W metodzie wektorów nośnych, mimo wykorzystania jedynie dwóch zmiennych, uzyskano lepsze wyniki niż w przypadku drzew decyzyjnych oraz metod uczenia zespołowego.

Dla SVM z jądrem radialnym najlepszy rezultat osiągnięto po optymalizacji parametrów

gamma i kosztu C — skuteczność wyniosła wtedy 55,6%. Natomiast w przypadku jądra liniowego, wyniki były niższe nawet od domyślnych ustawień jądra radialnego. Maksymalna skuteczność dla jądra liniowego wyniosła 51,4%, podczas gdy przy domyślnym parametrze gamma i zmiennym C w jądrze radialnym osiągnięto 54,8%.

Wśród metod uczenia zespołowego i drzew decyzyjnych, najlepsze rezultaty uzyskano stosując bagging, który pozwolił na zmniejszenie błędu klasyfikacji do około 20%. Metoda ta okazała się nieznacznie lepsza od lasów losowych (random forest).

Metoda SVM z odpowiednio dobranym jądrem i parametrami przewyższyła pozostałe podejścia pod względem skuteczności. W przypadku metod zespołowych, bagging uzyskał najniższy błąd, choć różnice były niewielkie.

2 Analiza skupień – algorytmy grupujące i hierarchiczne

2.1 Wybór i przygotowanie danych

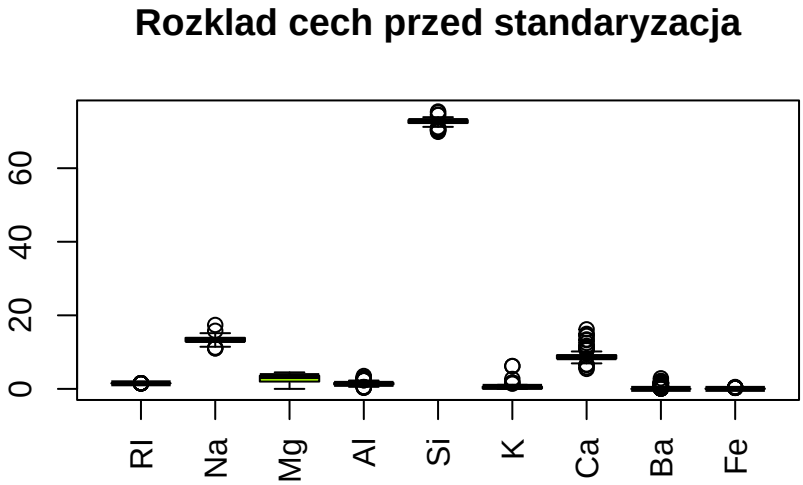
W tym zadaniu pracujemy na danych Glass (R-pakiet mlbench). Zbiór danych Glass charakteryzuje się złożoną strukturą klas oraz wyraźnymi różnicami w rozkładach cech chemicznych. W przeciwieństwie do prostych zbiorów jak iris, tutaj efekt maskowania klas jest silniejszy, a nierównowaga klas wymaga specjalnego podejścia.

Zbiór danych Glass zawiera 214 przypadków opisujących różne rodzaje szkła na podstawie ich składu chemicznego. Każdy przypadek charakteryzuje się 9 cechami numerycznymi, w tym zawartością pierwiastków takich jak sód (Na), magnez (Mg), glin (Al), krzem (Si), potas (K), wapń (Ca), bar (Ba), żelazo (Fe) oraz współczynnikiem załamania światła (RI). Klasyfikacja odbywa się na podstawie zmiennej Type, która określa typ szkła i przyjmuje 6 różnych wartości (od 1 do 6).

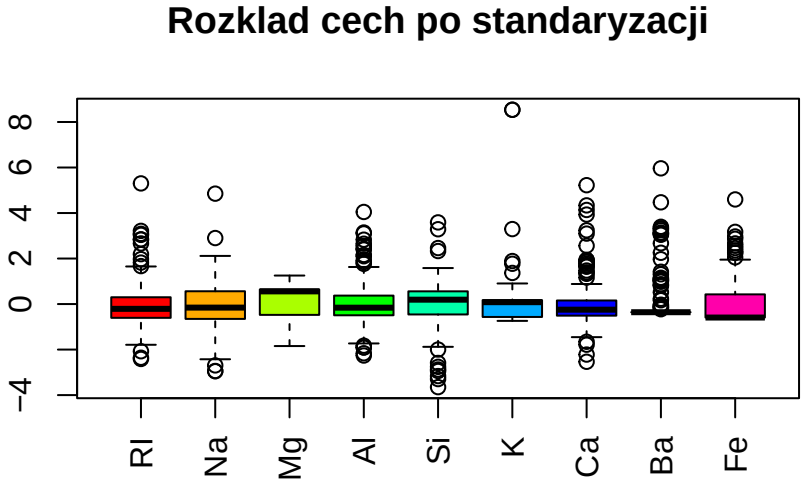
Nasze dane nie posiadają żadnych braków danych, brak występowania danego pierwiastka w danym rodzaju szkła oznaczany jest jako 0.0 zatem nie przeszkadza nam to w dalszej analizie.

Wszystkie typy zmiennych są numeryczne, z wyjątkiem zmiennej Type zawierającej etykietki naszych klas, która jest zmienną typu factor

Aby ułatwić wizualizację wyników, wybierzemy losowo (seed(123)) zbiór zawierający 200 rekordów (wierszy) oraz usuwamy zmienną grupującą zawierającą etykietki klas (grup).



Wykres pokazuje wyraźne różnice w skalach i rozproszeniu poszczególnych cech. W zbiorze danych Glass cechy reprezentują różne pierwiastki chemiczne (np. Na, Mg, Al) oraz współczynnik załamania światła (RI). Wartości tych cech mają różne zakresy. Jak widzimy na wykresie powyżej, pierwiastki mają różne zakresy, szczególnie zawartość krzemu jest zdecydownie wyższa niż innych pierwiastków. Może to znacząco wpłynąć na odległości, które zostaną zdominowane przez właśnie ten pierwiastek. Standaryzacja w tym przypadku jest zalecana tak aby obliczenia odległości nie zostały zdominowane przez cechy o większych wartościach, zaburzyłoby wyniki grupowania.

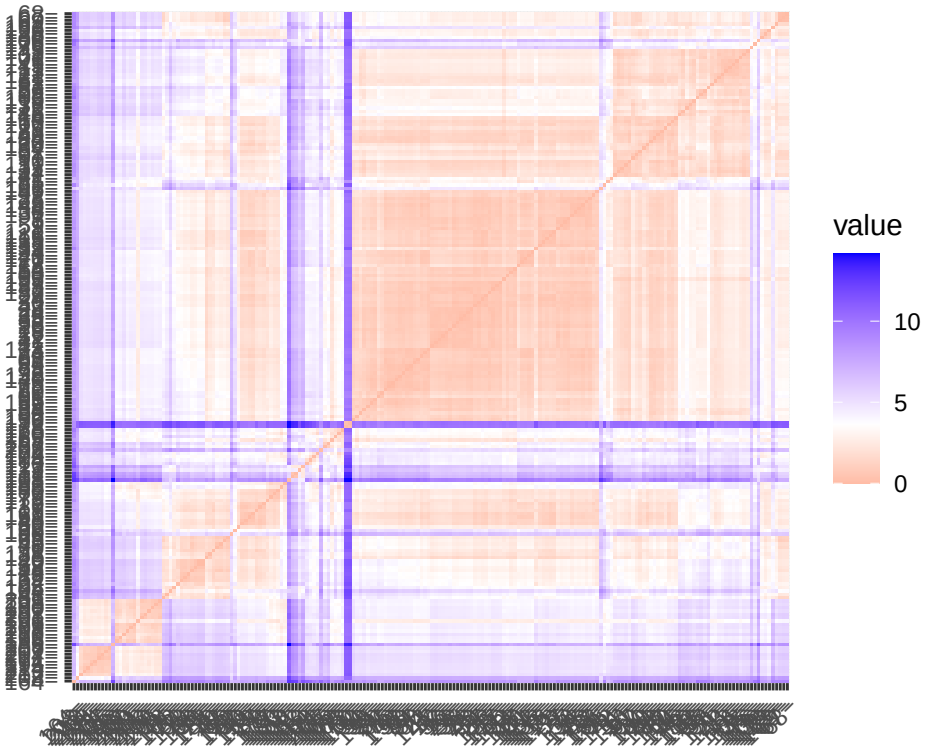


Na wykresie powyżej znajdują się rozkłady cech po standaryzacji. Jak widzimy, teraz wartości znajdują się w podobnym zakresie, dzięki czemu nie wpłyną fałszywie na ocene odległości podczas analizy skupień, którą zaraz wykonamy. ## Wizualizacja wyników algortymów

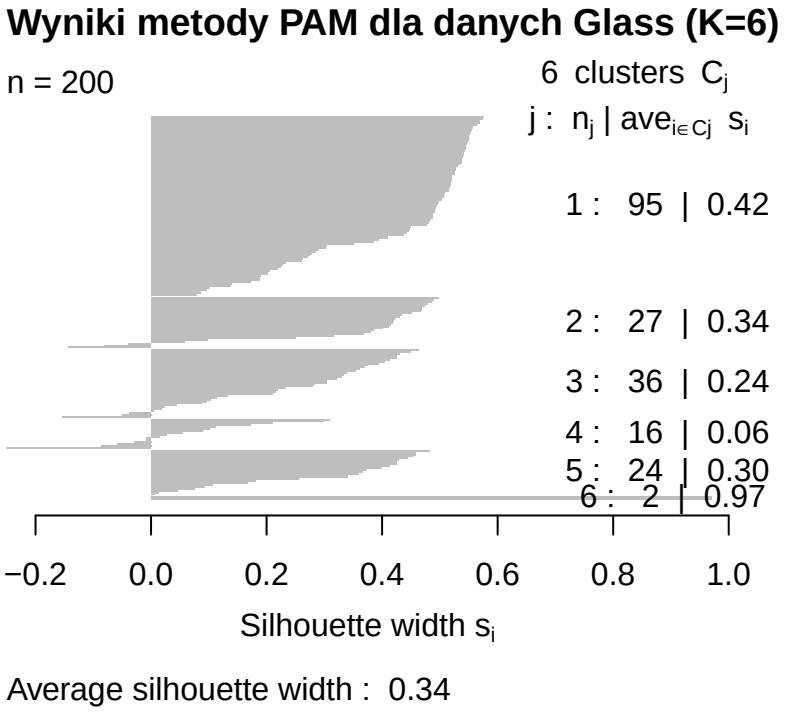
2.1.1 Zastosowanie algorytmów grupujących i hierarchicznych

W naszej analizie wezmiemy pod uwagę algorytm PAM (grupujący) oraz AGNES (hierarchiczny).

Na początku wyznaczymy macież niepodobieństw dla naszych danych.



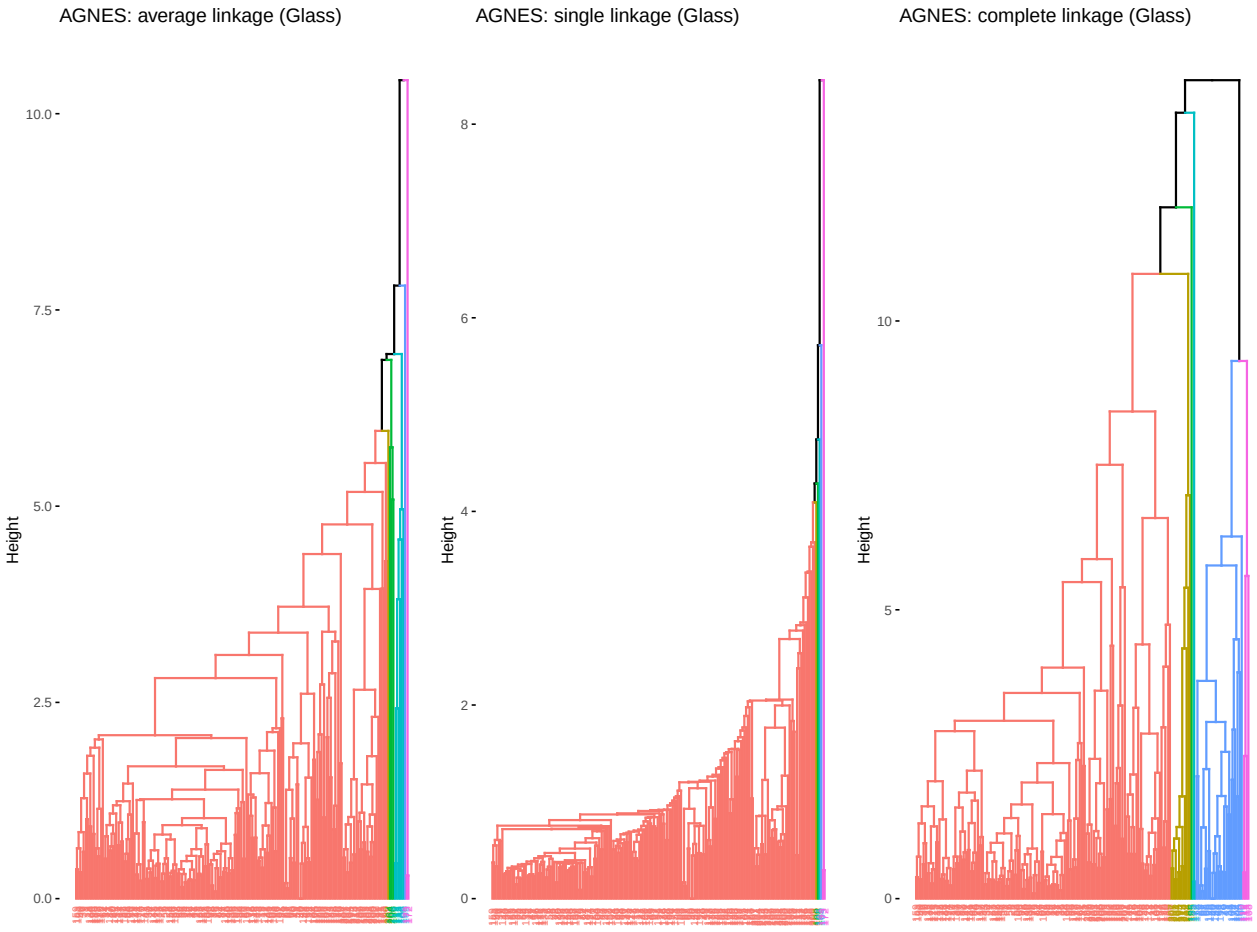
Na powyższym rysunku możemy zobaczyć macierz niepodobieństw dla naszych danych. Zastosujemy teraz algorytm PAM i zwizualizujemy jego wyniki. Przyjmujemy liczbę skupień K równą rzeczywistej liczbie klas, czyli w naszym przypadku 6.



Na powyższym wykresie widzimy wyniki algorymu PAM dla naszych danych. Widzimy podział na 6 skupisk, gdzie szerokość wynosi 0.34. Możemy zauważyć że największa ilość przypadków

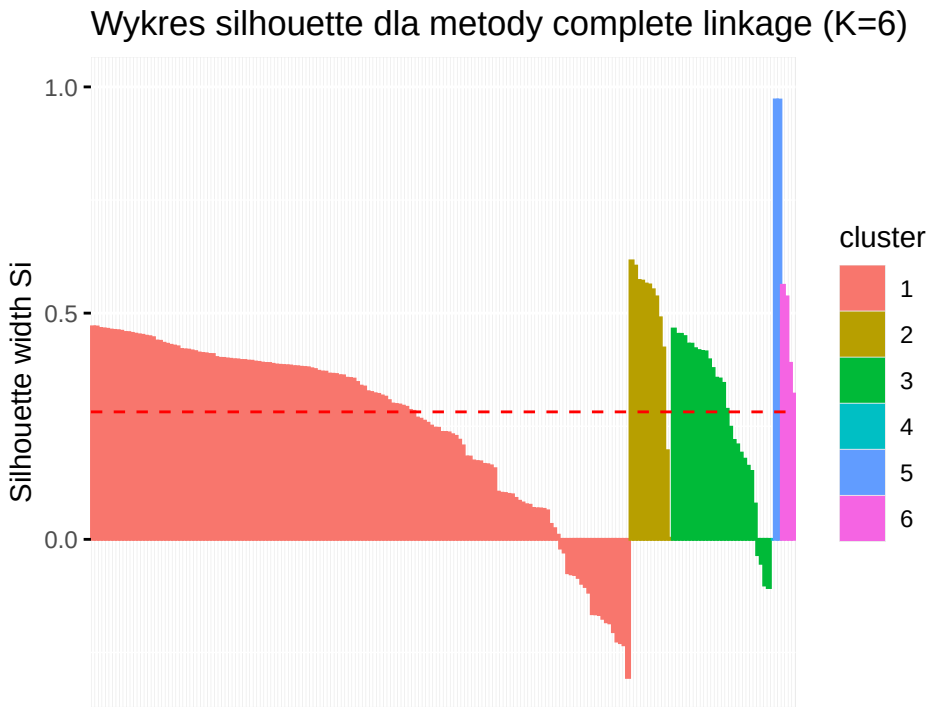
trafiła do klasy 1, natomiast najmniej do klasy 6, co może być zgodne z intuicją.

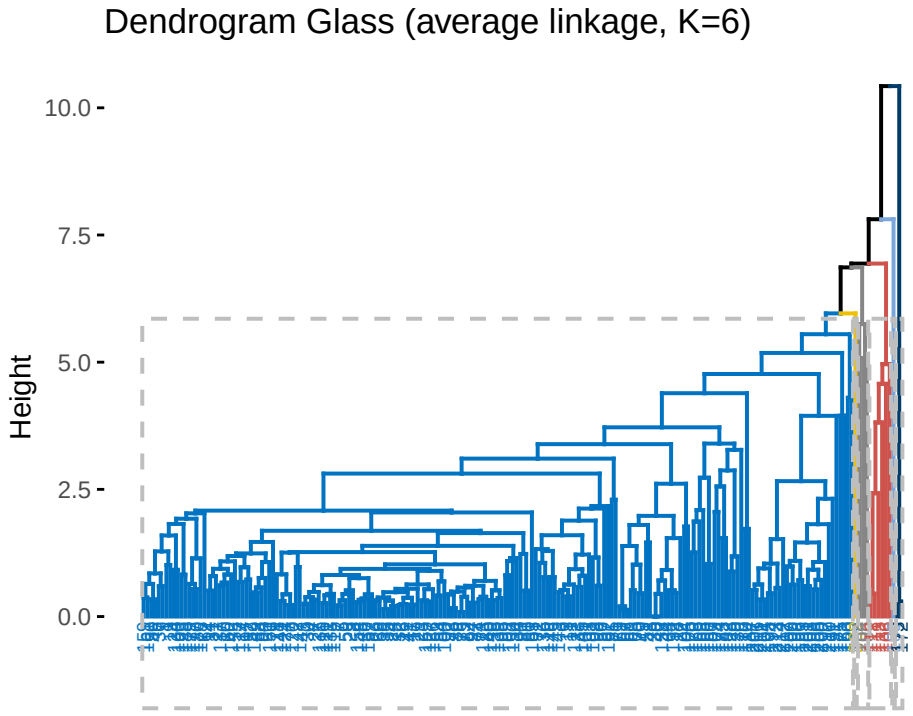
Spróbujmy teraz zastosować algorytm AGNES i zwizualizować jego wyniki. Przyjmujemy liczbę skupień K równą rzeczywistej liczbie klas, czyli w naszym przypadku 6.



Na powyższych wykresach widzimy nasze drzewa z podziałem na 6 skupisk. Widzimy, że najlepiej poradziła sobie metoda complete, co potwierdza analiza współczynnika aglomeracji, ktory wynosi dla niej: 0.9256946.

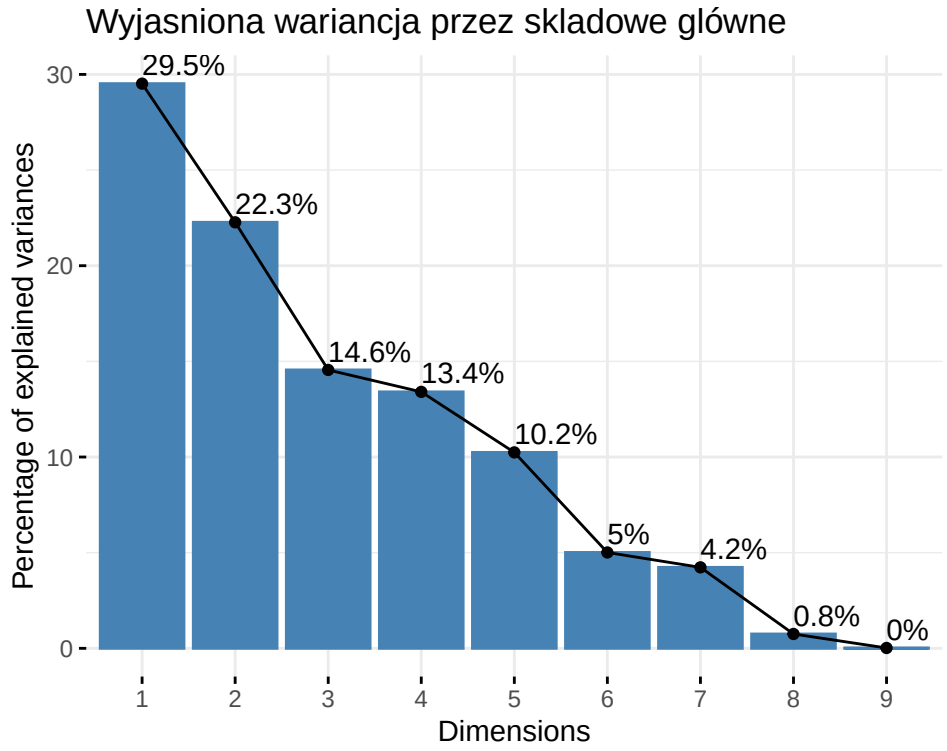
W dalszej analizie będziemy brali pod uwagę właśnie metodę complete. Narysujmy sobie jeszcze wykres wskaźnika silhouette dla k=6.



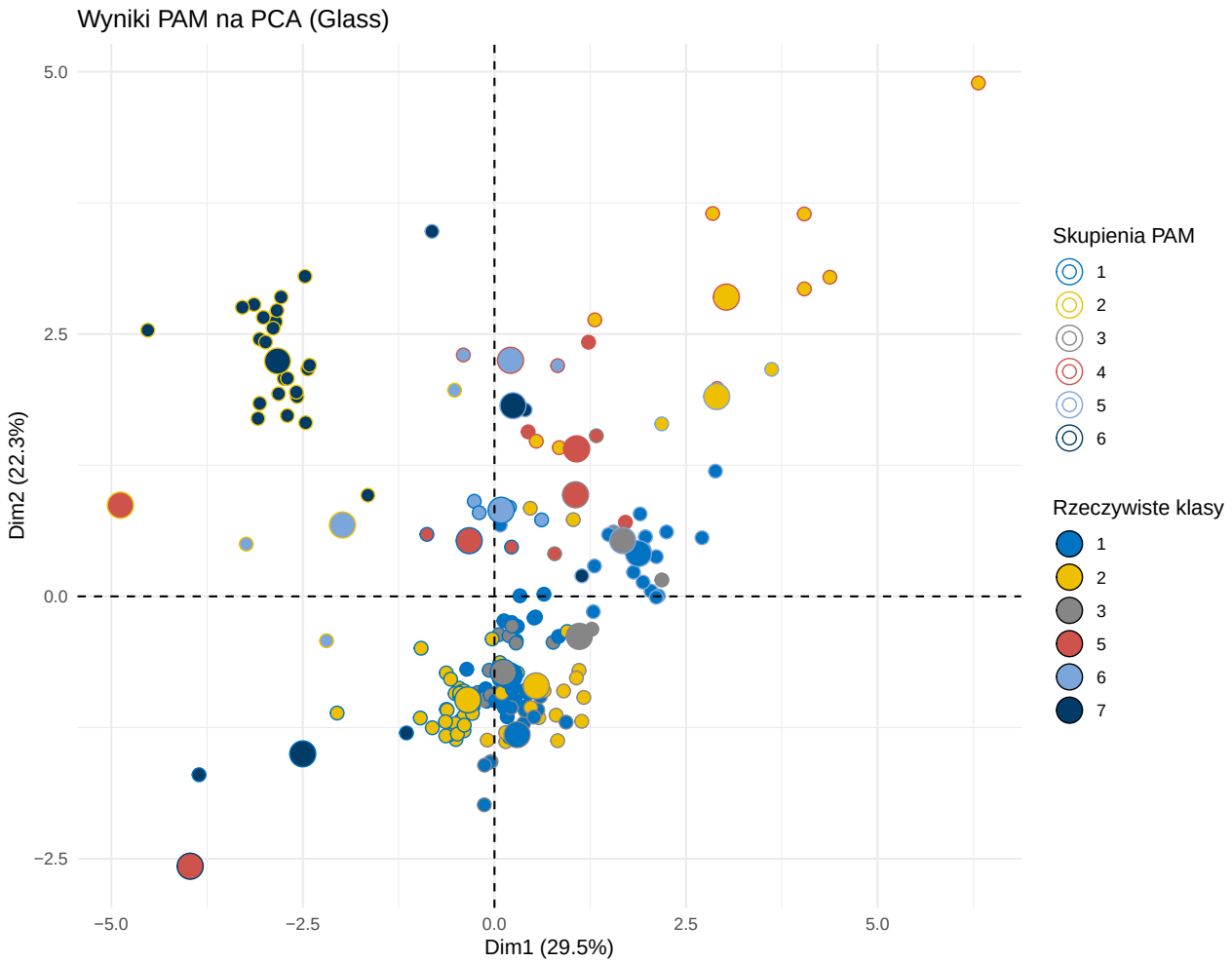


2.1.2 Wizualizacja danych na bazie PCA

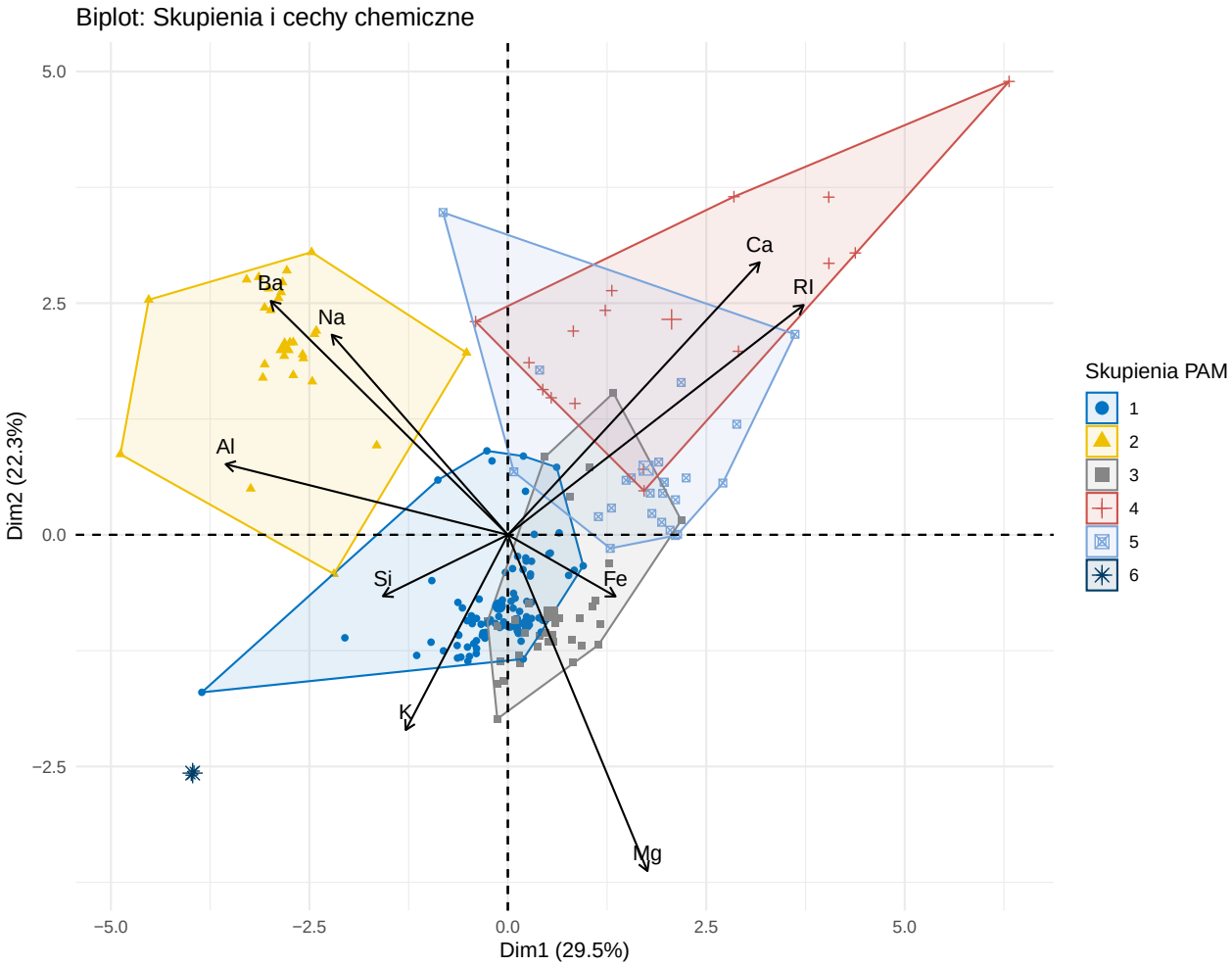
Zwizualizujmy teraz nasze dane na wykresie 2D. Wykorzystamy do tego metodę PCA. Na początku weźmiemy pod uwagę wyniki algorytmu grupującego - PAM.



Na powyższym wykresie możemy zobaczyć wariancję kolejnych składowych. Z dwóch pierwszych otrzymujemy jedynie ok. 52% pełnej informacji, natomiast dla czytelności wykresu, do zobrazowania użyjemy jedynie dwóch pierwszych składowych.

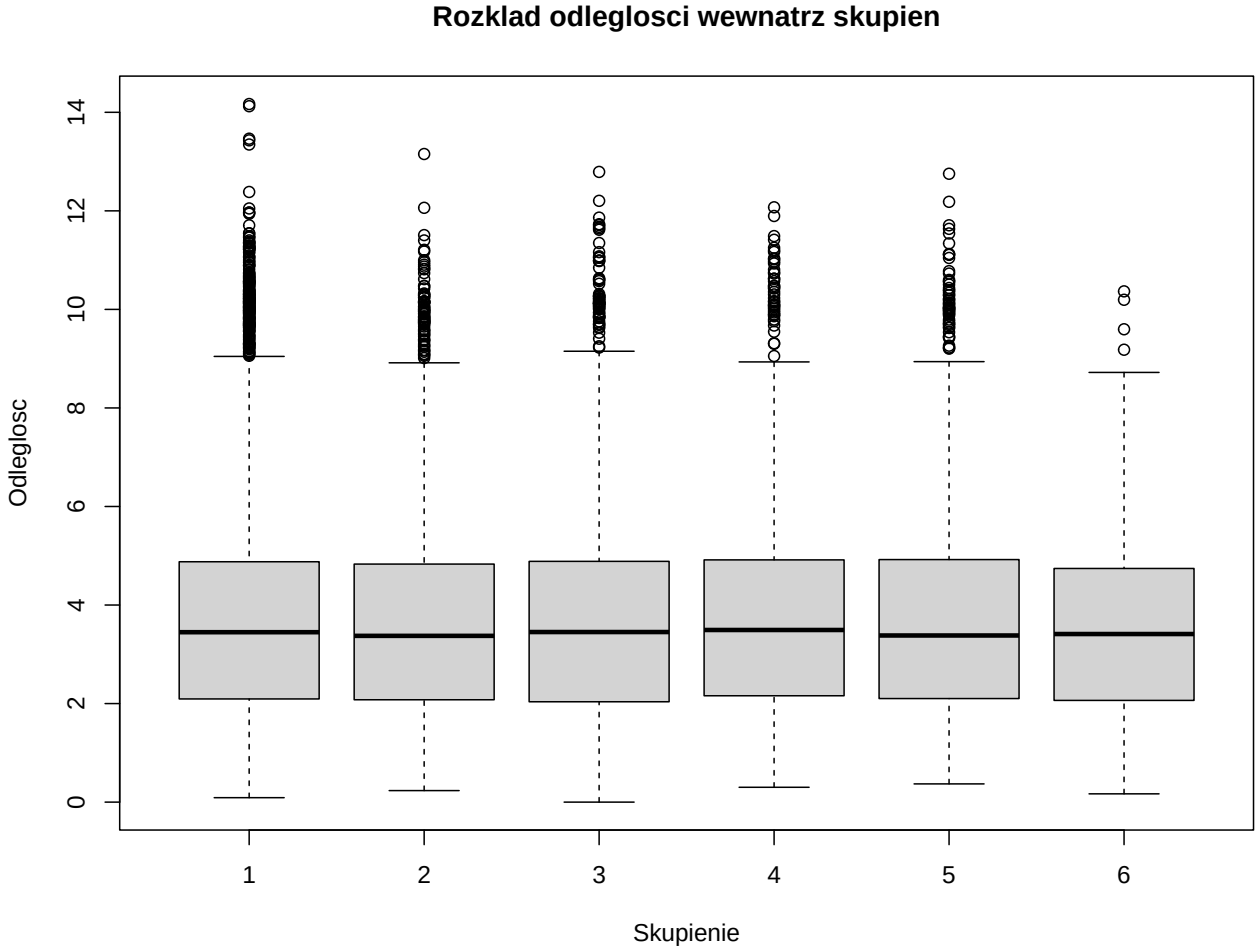


Na powyższym wykresie możemy zobaczyć Przynależność do skupisk i rzeczywistych typów naszego szkła. Nasz algorytm nie poradził sobie za dobrze z przyporządkowaniem skkupień do rzeczywistych klas. Widzimy obszary w których punkty z różnych skupisk się mieszają. Spróbujmy jeszcze zwizualizować nasze dane poprzez biblot:



Tutaj dokładnie widzimy jak wyglądają nasze skupienia punktów. Niektóre niestety nachodzą na siebie, natomias biorąc pod uwagę dużą ilość klas, nasz wykres nie wygląda najgorzej. Możemy jeszcze spojrzeć na centra klas, jednak w naszym przypadku nie wniosą one za dużo do naszej analizy.

Popatrzmy teraz na własności naszych skupień:



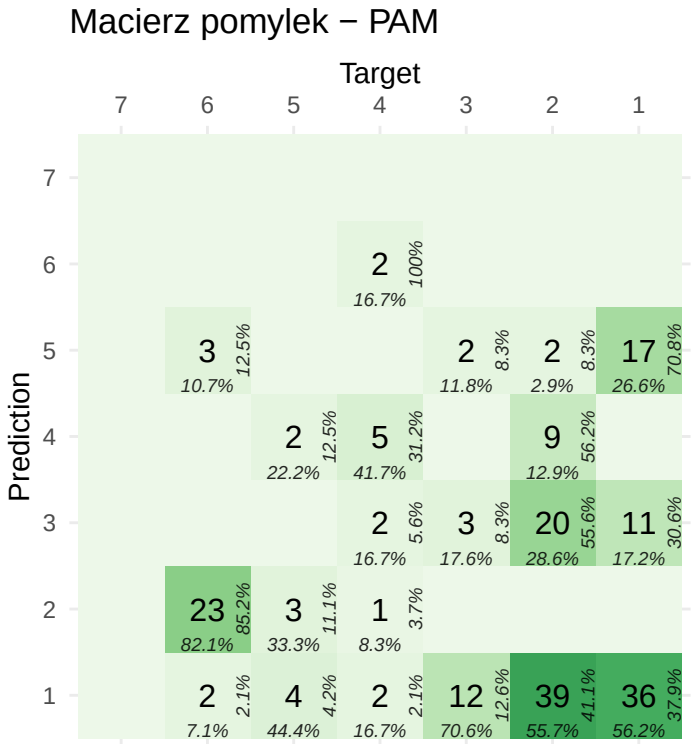
Na powyższym wykresie możemy zobaczyć rozkłady odległości wewnątrz każdego z naszych skupień. Widzimy że zarówno wariancje jak i średnie są do siebie zbliżone. Świadczy to o podobnym poziomie „zwartości" wewnątrz skupień. Odstające obserwacje mogą sugerować, że część punktów jest słabo dopasowana do przypisanego skupienia.

Średnia odległość wewnątrz skupień wynosi: 2.2079758 co wskazuje na umiarkowaną zwartość grup.

Średnia odległość między skupieniami wynosi 4.4263921 co wskazuje na ogólną umiarkowaną separację, natomiast wskaźnik separacji wynosi: 0.5723159, 1.3437414, 0.5723159, 1.0270312, 0.8235244, 8.4544305.

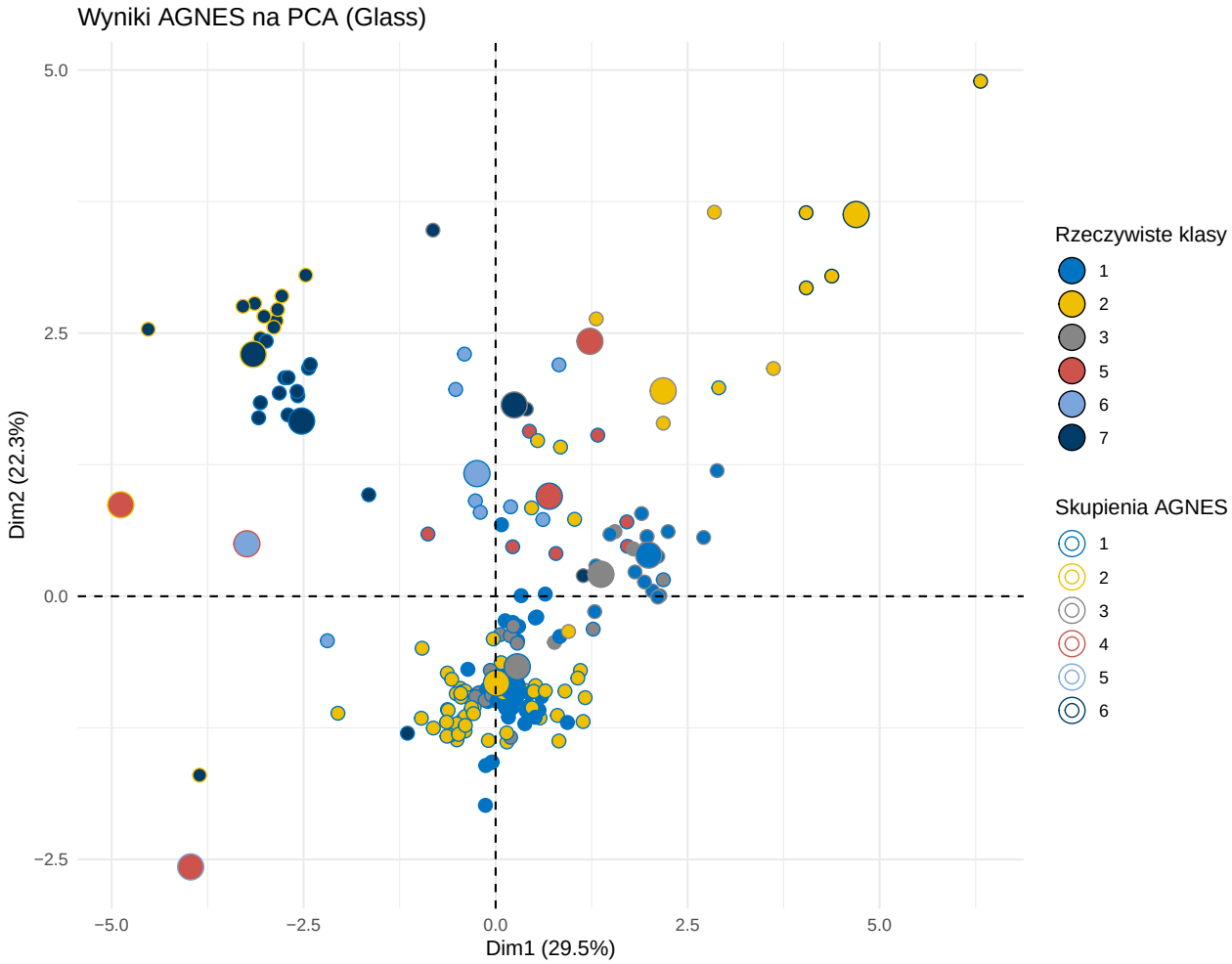
Poszczególne pary skupień wykazują jednak zróżnicowaną separację: Ponad połowa par ma słabą lub umiarkowaną separację (wartości 0.57-1.34) Jedna para skupień jest wyraźnie oddzielona (wartość 8.45)

Wyznaczmy teraz macierz pomyłek, tak aby zobaczyć separowalność poszczególnych klas.

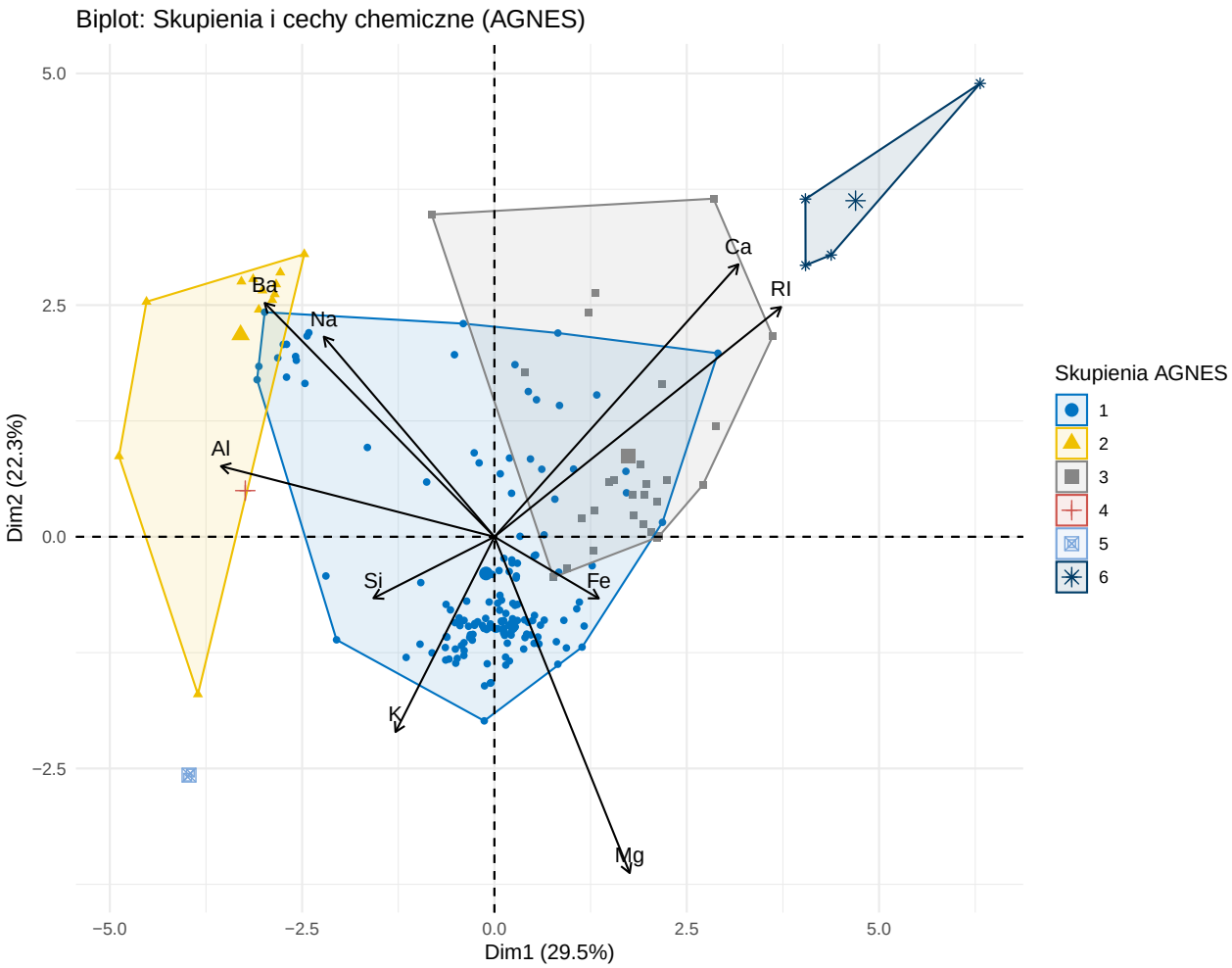


Widzimy że nasz algorytm PAM poradził sobie słabo z rozpoznawaniem typów, zgodność tutaj wynosiła zaledwie 0.22.

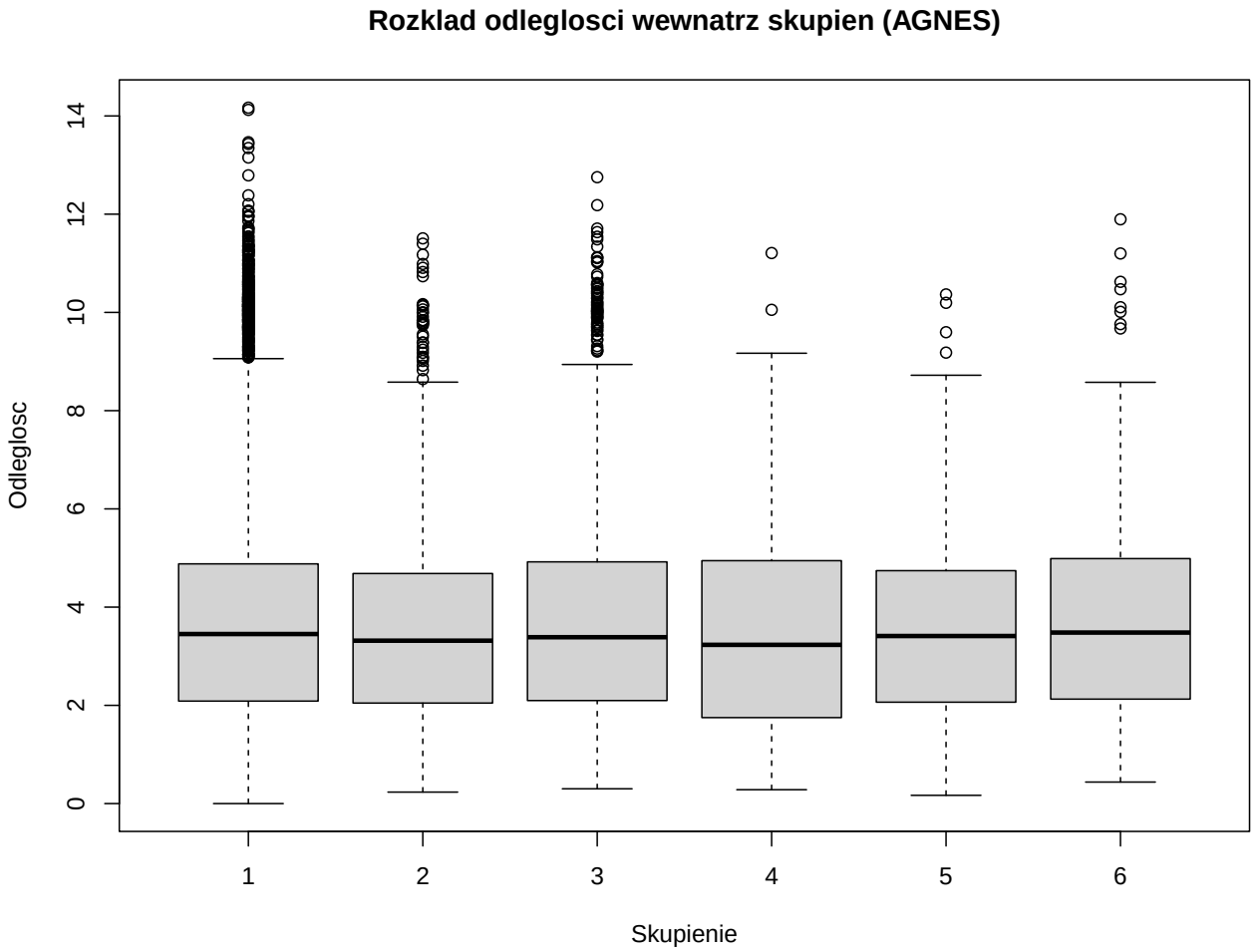
Przeanalizujmy teraz wyniki algorytmu hierarchicznego - AGNES. Wariancja w naszym przypadku pozostaje bez zmian, używamy tego samego PCA co w poprzeniej wizualizacji.



Na powyższym wykresie możemy zobaczyć Przynależność do skupisk i rzeczywistych typów naszego szkła. Agnes również nie poradził sobie za dobrze z przyporządkowaniem skkupień do rzeczywistych klas. Widzimy obszary w których punkty z różnych skupisk się mieszają. Spróbujmy jeszcze zwizualizować nasze dane poprzez biblot:



Biplot również wskazuje na słabe rozdzielenie i nakładanie się kolejnych skupień.

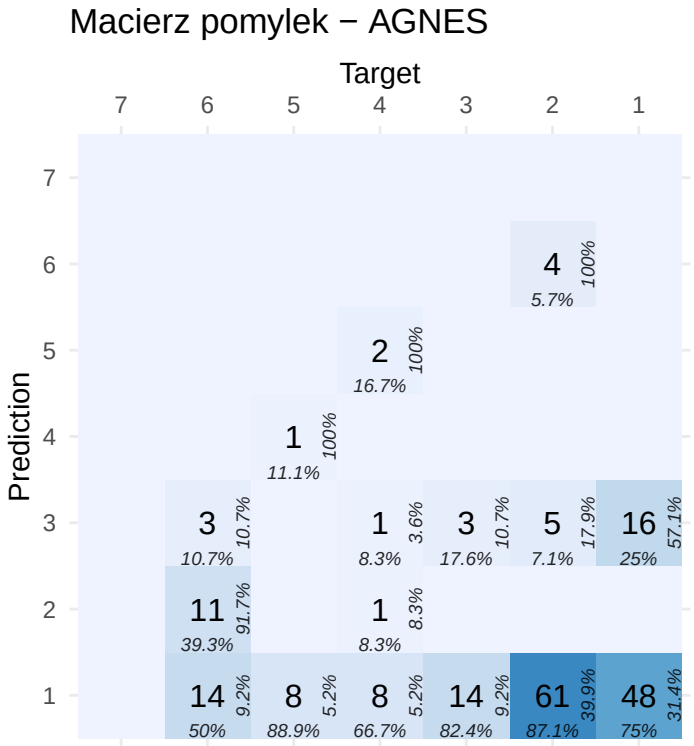


Na powyższym wykresie możemy zobaczyć rozkłady odległości wewnątrz każdego z naszych skupień. Widzimy że tutaj również wariancje jak i średnie są do siebie zbliżone.

Średnia odległość wewnątrz skupień wynosi: 2.7387646 co wskazuje na umiarkowaną zwartość grup.

Średnia odległość między skupieniami wynosi 4.4263921 co wskazuje na ogólną umiarkowaną separację, natomiast wskaźnik separacji wynosi: 0.5723159, 1.3437414, 0.5723159, 1.0270312, 0.8235244, 8.4544305.

Wyznaczmy teraz macierz pomyłek, tak aby zobaczyć separowalność poszczególnych klas.

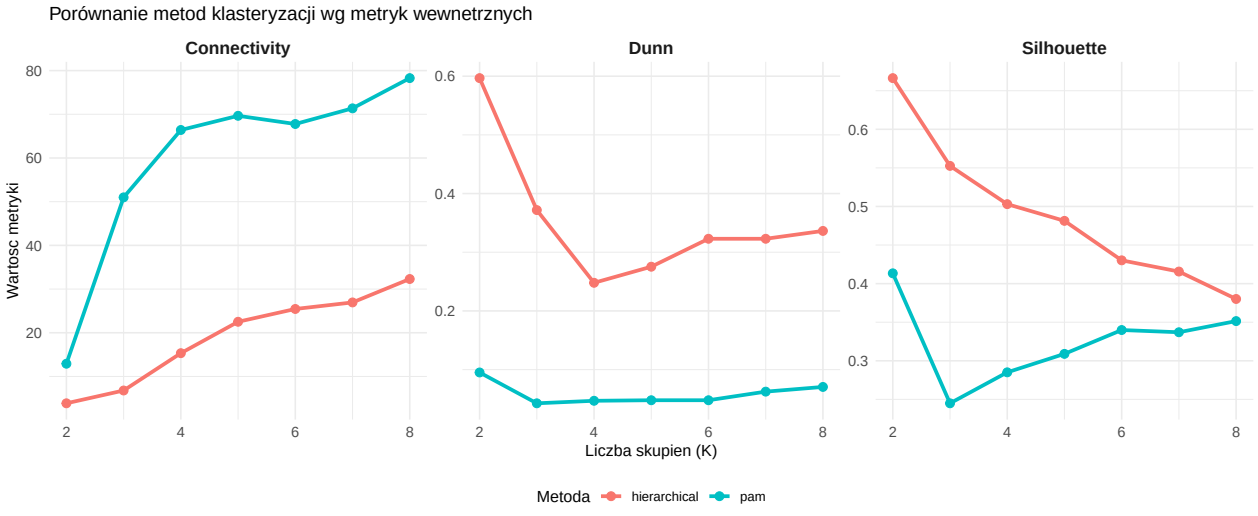


Widzimy że nasz algorytm AGNES lepiej sobie poradził z rozpoznawaniem typów, poniważ zgodność tutaj wynosiła 0.255. Jednak nie jest to w dalszym ciągu dobry wynik.

2.2 Ocena jakości grupowania. Wybór optymalnej liczby skupień i porównanie metod.

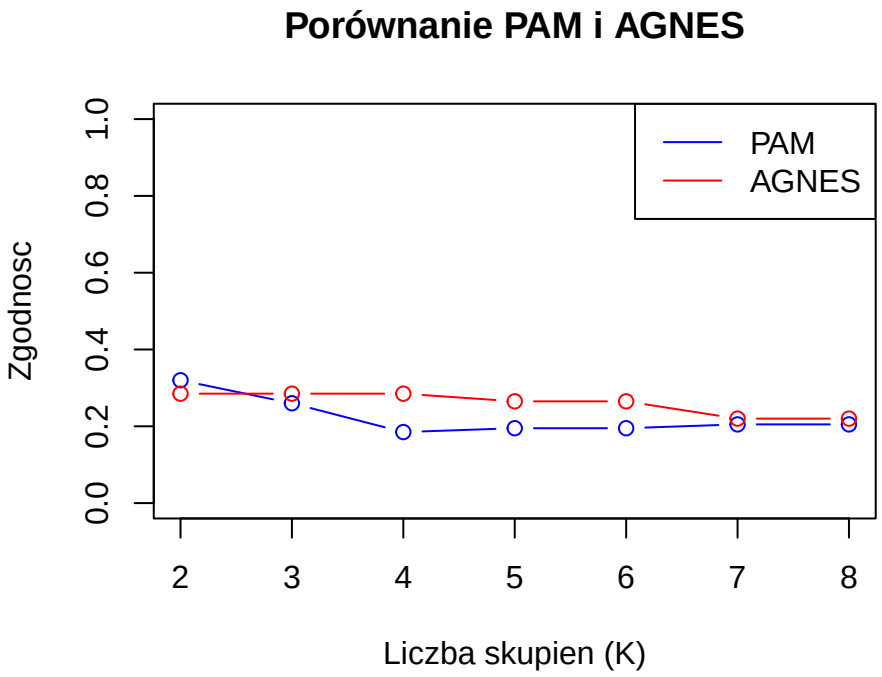
2.2.1 Wskaźniki wewnętrzne

Do oceny wskaźników wewnętrznych wykorzystamy średnią wartość indeksu silhouette, dunn oraz conectivity do porównania wyników otrzymanych dla różnych algorytmów analizy skupień (PAM i AGNES) oraz różnej liczby skupień K.



Analiza wskaźników wewnętrznych dla metod hierarchicznej (AGNES) i PAM wykazuje wyraźne różnice w jakości grupowania. Dla metody hierarchicznej obserwujemy lepsze wyniki - niską wartość Connectivity (3.86 dla K=2), wysokie Silhouette (0.67 dla K=2) i przyzwoity indeks Dunna (0.60 dla K=2), co wskazuje na dobre zwartość i separację skupień. Jakość stopniowo spada wraz ze wzrostem liczby skupień. Metoda PAM wypadła gorzej - wysokie Connectivity (12.9 dla K=2) i niskie wartości Dunna (<0.1) sugerują problemy z separacją grup. Dla obu metod optymalne wydaje się K=2, gdzie hierarchiczna osiąga najlepsze wyniki we wszystkich metrykach. Wniosek: metoda hierarchiczna z K=2 to najlepszy wybór dla tych danych.

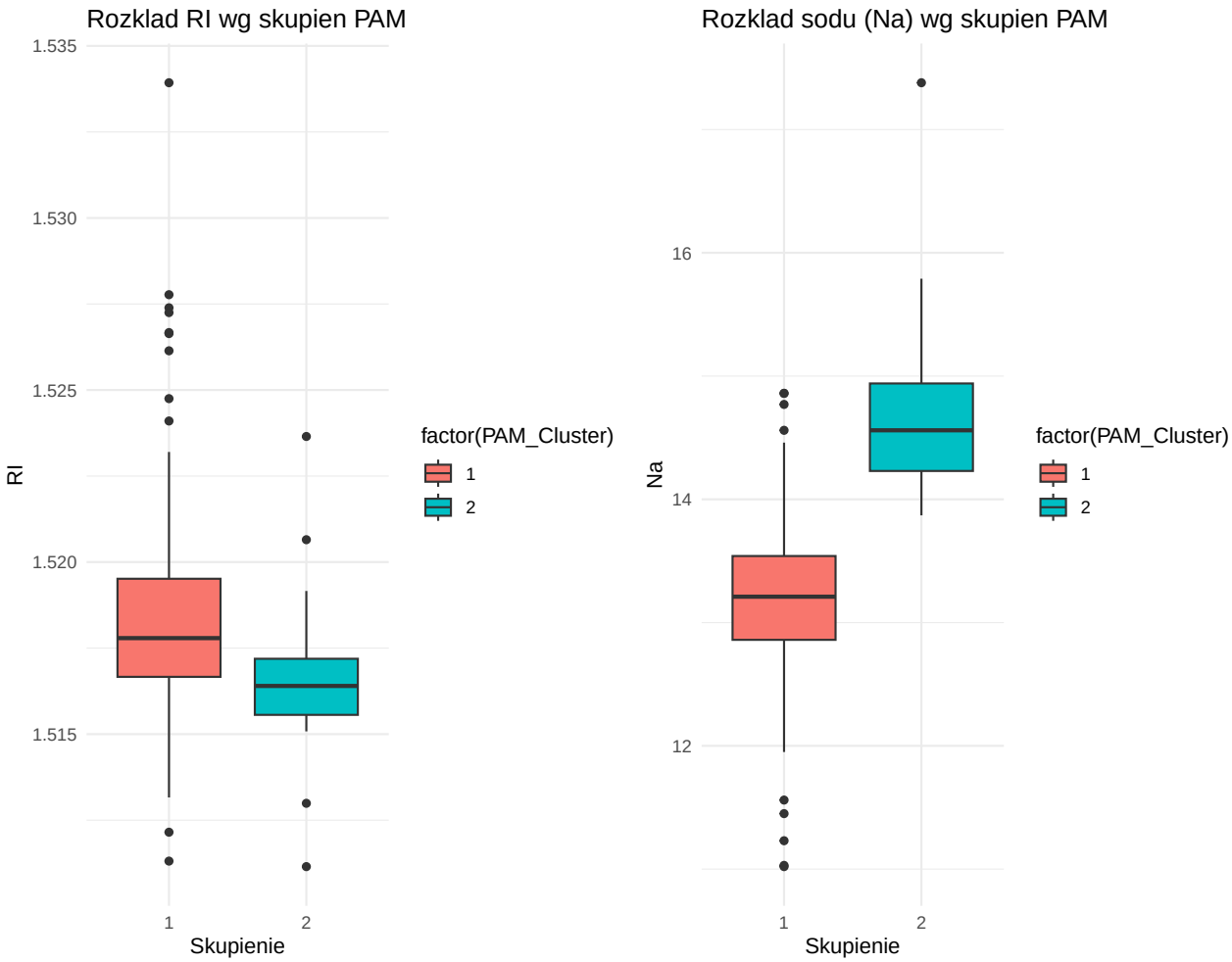
2.2.2 Wskaźniki zewnętrzne



Analiza wyników pokazuje, że zarówno metoda PAM, jak i AGNES osiągają najwyższą zgodność z rzeczywistymi klasami (ok. 32%) dla K=2, przy czym wyniki PAM są nieco lepsze w tym przypadku. Dla większych wartości K obserwujemy stopniowy spadek zgodności, przy czym AGNES konsekwentnie uzyskuje lepsze wyniki niż PAM - dla K=3 PAM osiąga 26% vs 28,5% AGNES, a dla K=8 różnica wynosi 20.5% vs 22%. Wartości dla PAM K>4 utrzymują się na stabilnym, ale niskim poziomie , co sugeruje, że podział na więcej niż 2-3 skupienia pogarsza zgodność z rzeczywistą strukturą klas w danych. Optymalnym wyborem jest zatem K=2 dla obu metod.

2.3 Interpretacja wyników grupowania - charakterystyki skupień

Wyznaczamy podział na skupienia dla optymalnej liczby skupień K, czyli w naszym przypadku 2. Popatrzmy najpierw na wykresy pudełkowe dwóch wybranych cech:



Przedstawione boxploty pokazują, że klasteryzacja PAM skutecznie rozdzieliła dane na dwa skupienia w oparciu o zmienne RI i Na; skupienie 1 ma nieco wyższe RI i znacznie niższe Na, podczas gdy skupienie 2 charakteryzuje się niższym RI i wyraźnie wyższym Na, co sugeruje, że zawartość sodu jest kluczowym czynnikiem różnicującym te grupy.

Popatrzmy teraz na dane medioidów:

Tabela 1: Porownanie medoidow i srednich

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
Medoid 1	1.52	13.21	3.39	1.33	72.76	0.59	8.59	0.00	0.00
Średnia 1	1.52	13.21	3.08	1.32	72.60	0.56	9.01	0.02	0.07
Medoid 2	1.52	14.95	0.00	2.27	73.30	0.00	8.71	0.67	0.00
Średnia 2	1.52	14.67	0.30	2.15	73.08	0.13	8.57	1.02	0.01

Medoidy w metodzie PAM są dobrane tak, by reprezentowały średnie charakterystyki swoich skupień. Analiza pokazuje, że Medoid 1 ściśle odpowiada średnim wartościom swojego skupienia, szczególnie w zawartości sodu (Na: 13.21) i krzemu (Si: ~72.7), potwierdzając jego

typowość. Z kolei Medoid 2, choć ogólnie odzwierciedla wysokosodowy profil skupienia (Na: 14.95), wykazuje pewne ekstremalne cechy – całkowity brak magnezu (Mg: 0.00) i potasu (K: 0.00) oraz niższą zawartość baru (Ba: 0.67) niż średnia skupienia. Różnice te podkreślają główne kontrasty między skupieniami: Skupienie 1 to próbki z magnezem i umiarkowanym sodem, podczas gdy Skupienie 2 to wysokosodowe próbki z barem, ale pozbawione magnezu.

Spójrzmy teraz na wykres PCA

Analiza PCA dla K=2 pokazuje wyraźne różnice w rozmieszczeniu skupień między metodami PAM i AGNES. W przypadku PAM, lewa górna ćwiartka wykresu jest zdominowana przez skupienie 2, podczas gdy pozostałe obszary zawierają głównie skupienie 1, co wskazuje na dobrą separację między tymi grupami. Z kolei w metodzie AGNES prawa górna ćwiartka skupia większość obiektów skupienia 1, a reszta wykresu wypełniona jest głównie obiektami skupienia 2, co sugeruje nieco inną organizację danych. Rozkład ten potwierdza, że PAM lepiej oddziela skupienia w przestrzeni PCA, podczas gdy AGNES tworzy bardziej zrównoważone, ale mniej wyraźnie rozdzielone grupy. Różnice te wynikają z odmiennych zasad działania obu metod – PAM skupia się na reprezentantach (medoidach), podczas gdy AGNES łączy obiekty hierarchicznie, co może prowadzić do łagodniejszych granic między skupieniami.

Podsumowanie: Analiza średnich wartości cech i wykresów pudełkowych ujawniła wyraźne różnice między skupieniami. Skupienie 1 charakteryzuje się niższą średnią zawartością sodu (Na: ~13.2) i obecnością magnezu (Mg: ~3.1), co wskazuje na próbki o bardziej zrównoważonym składzie chemicznym. W Skupieniu 2 dominuje wysoka zawartość sodu (Na: ~14.7) i wyraźna obecność baru (Ba: ~1.0), przy niemal zerowym poziomie magnezu (Mg: ~0.3), co sugeruje próbki o specyficznym, wyspecjalizowanym składzie. Wykresy pudełkowe potwierdzają te różnice, pokazując minimalne nakładanie się rozkładów kluczowych cech (np. Na, Mg) między skupieniami. Rozbieżności w wartościach medoidów względem średnich (zwłaszcza w Skupieniu 2) wskazują na istnienie podgrup wewnątrz skupień, co może wymagać dalszej analizy dla lepszego zrozumienia struktury danych.