

Raport Lista 4

Eksploracja danych

Dominik Kowalczyk i Matylda Mordal

2025-06-18

Spis treści

1	Zaawansowane metody klasyfikacji	1
1.1	Opis danych	1
1.2	Rodziny klasyfikatorów/uczenie zespołowe (ensemble learning)	3
1.3	Metoda wektorów nośnych (SVM)	10
1.4	Porównanie skuteczności metod	16
2	Analiza skupień - algorytmy grupujące i hierarchiczne	16
2.1	Przygotowanie danych	16
2.2	Wizualizacja wyników grupowania	18
2.3	Ocena jakości grupowania. Wybór optymalnej liczby skupień i porównanie metod.	24
2.4	Interpretacja wyników grupowania - charakterystyki skupień	27

1 Zaawansowane metody klasyfikacji

1.1 Opis danych

Do wykonania zadania wykorzystamy zbiór danych Glass (mlbench). Opisuje on dane identyfikacyjne szkła używane oraz potrzebne przy śledztwach kryminalistycznych. Przyjrzymy się zatem, co znajduje się w analizowanym zbiorze.

Tabela 1: Opis danych Glass

Indeks	Nazwa zmiennej	Typ zmiennej	Opis zmiennej
1	RI	numeric	Współczynnik załamania światła
2	Na	numeric	Zawartość sodu
3	Mg	numeric	Zawartość magnezu
4	Al	numeric	Zawartość glinu
5	Si	numeric	Zawartość krzemu

6	K	numeric	Zawartość potasu
7	Ca	numeric	Zawartość wapnia
8	Ba	numeric	Zawartość baru
9	Fe	numeric	Zawartość żelaza
10	Type	factor	Typ szkła

Liczba przypadków w zbiorze danych `glass_data` wynosi 214.

Zbiór danych `glass_data` zawiera 10 zmiennych, z czego ostatnia z nich, `Type` przechowuje informacje o przynależności obiektu do konkretnej klasy (tzw. etykieta klas). Jest ona typu: factor. Dodatkowo, pozostałe zmienne zawierają informację dotyczące występowania danego pierwiastka chemicznego w szkle i są one typu numeric, co pozwala nam stwierdzić, że wszystkie zmienne w naszym analizowanym zbiorze mają prawidłowo przypisane typy.

Tabela 2: Liczba Obserwacji dla Każdego Typu Szkła

Typ.Szkła	Liczba.Obserwacji
1	70
2	76
3	17
5	13
6	9
7	29

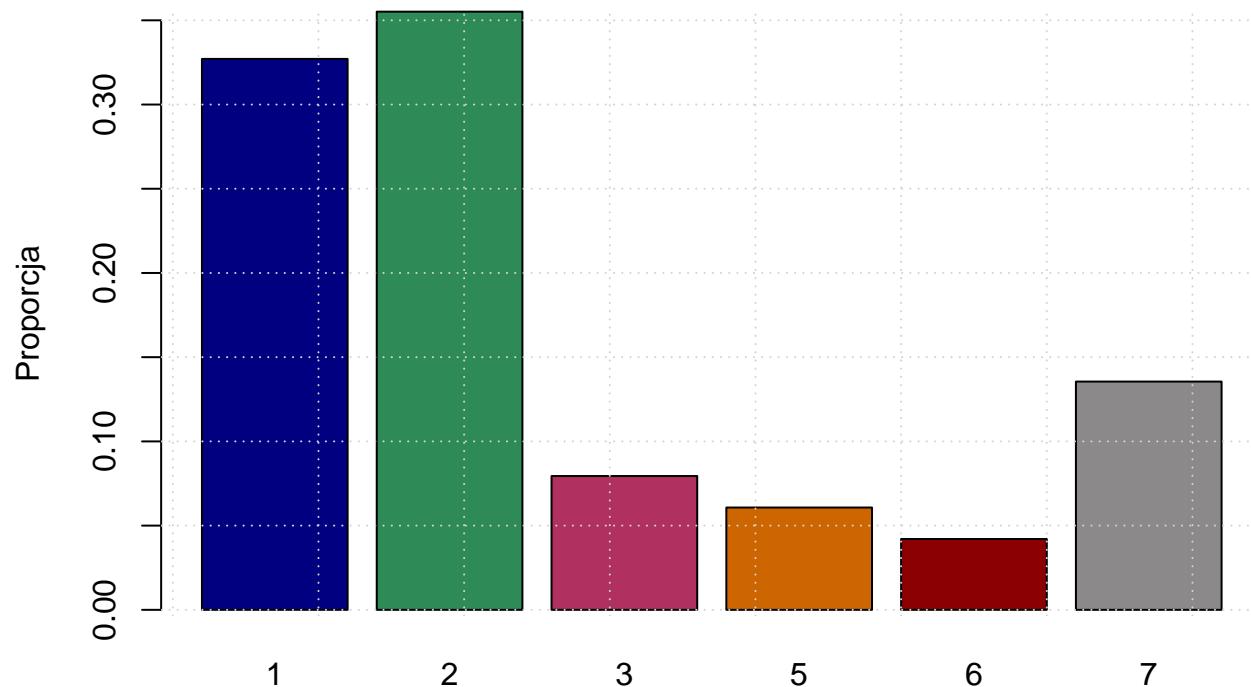
```
#Czy istnieją jakieś braki w danych?
any(is.na(glass_data)) ||
any(sapply(glass_data, function(col) is.character(col) & (col == "" | grep(" ", col))))
```

[1] FALSE - Zatem nasz zbiór danych jest kompletny i nie występują tam żadne braki danych. Sprawdźmy rozkład danych w celu zrozumienia z czym mamy doczynienia, jak również poszukując nieścisłości lub różnego rodzaju nietypowych wartości.

W analizowanym zbiorze nie mamy doczynienia z “nieścisłościami” rozkładów w sensie błędów w danych, ale raczej charakterystycznymi cechami chemicznymi różnych typów szkła. Dziwne rozkłady (silnie skośne, z licznymi odstającymi) dla takich pierwiastków jak Mg, K i Ba są prawdopodobnie wynikiem specyficznych receptur chemicznych stosowanych do wytwarzania różnych rodzajów szkła o odmiennych właściwościach (np. szkło budowlane vs. szkło optyczne vs. szkło kryształowe), szczególnie że ilość obserwacji dla każdego typu szkła znacznie się różni (dla klasy 1 jest 70 a dla 6 tylko 9).

1.2 Rodziny klasyfikatorów/uczenie zespołowe (ensemble learning)

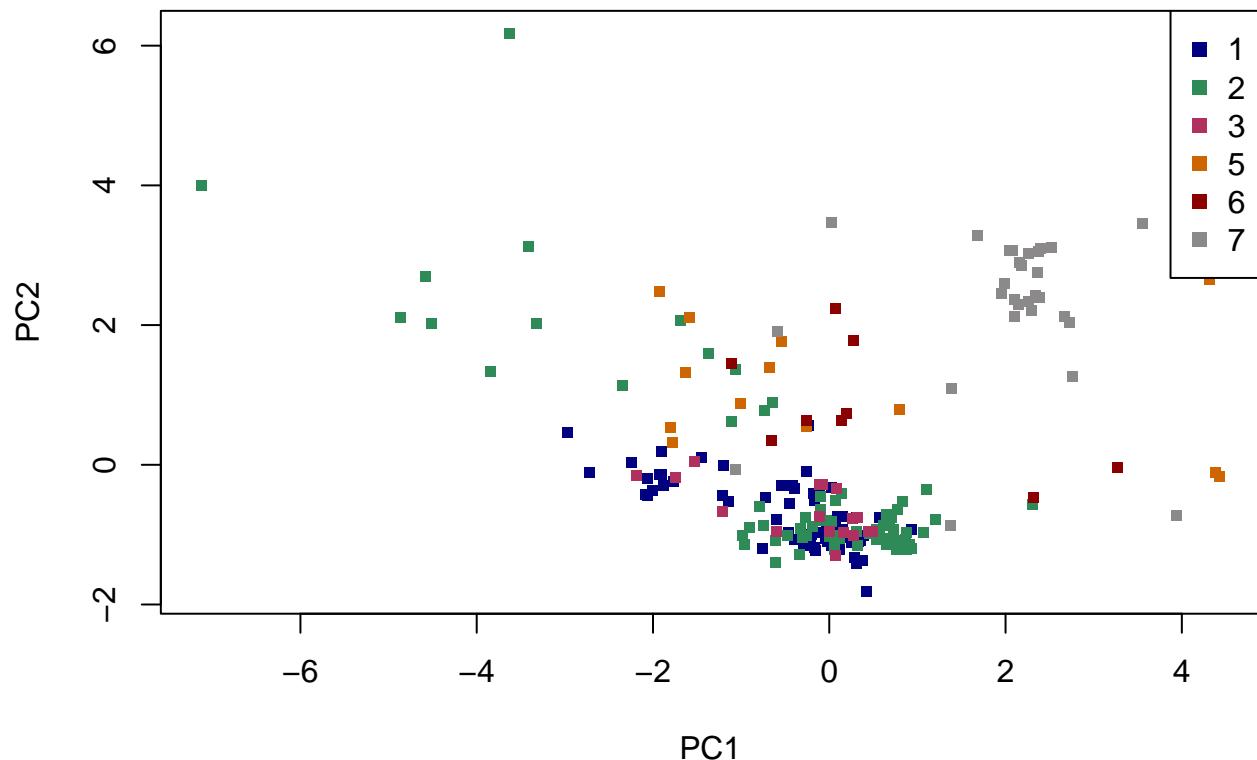
Dane Glass – rozkład klas



Rysunek 1: Rozkład klas w zbiorze danych Glass

Rysunek 1 obrazuje, jak rozkładają się klasy w zbiorze danych **Glass**. Na osi poziomej mamy numery klas od 1 do 7, a na osi pionowej widzimy proporcje, które sięgają ponad 0,30. Najwięcej danych przypada na klasę 2, a następna pod względem wielkości jest klasa 1. Klasy 3, 5, 6 i 7 mają znacznie mniejsze proporcje, wszystkie poniżej 0,15.

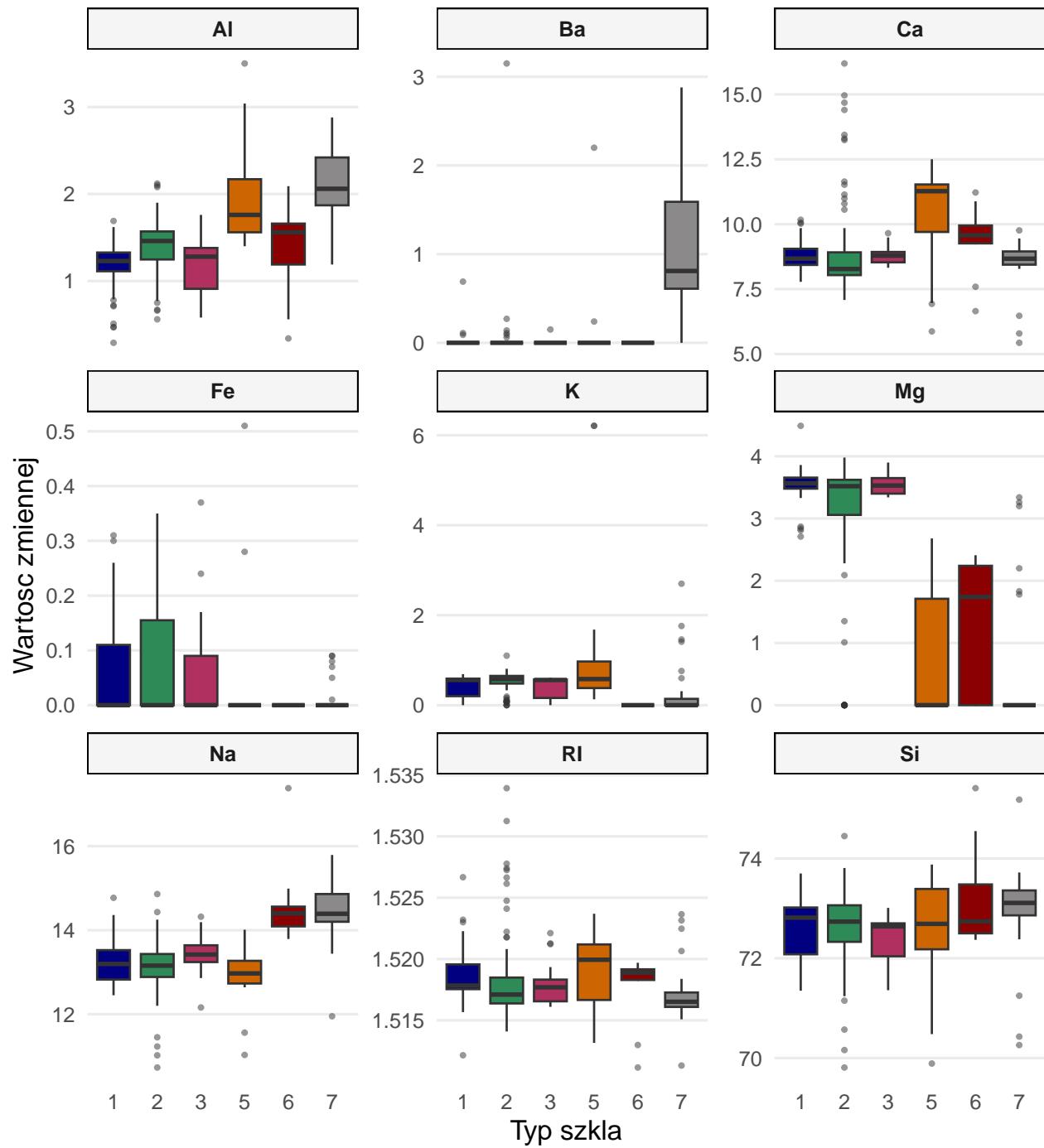
Dane Glass – PCA



Rysunek 2: PCA dla danych Glass

Z kolei Rysunek 2, czyli PCA dla danych Glass, wizualizuje dane po zastosowaniu analizy głównych składowych. Wykres ten prezentuje punkty danych w dwuwymiarowej przestrzeni, zdefiniowanej przez dwie pierwsze składowe główne (PC1 i PC2). Na podstawie tej wizualizacji można zauważyc, że klasa 7 jest stosunkowo dobrze oddzielona od pozostałych, natomiast inne klasy w znacznym stopniu na siebie nachodzą, co sugeruje trudności w ich liniowej separacji w tej przestrzeni.

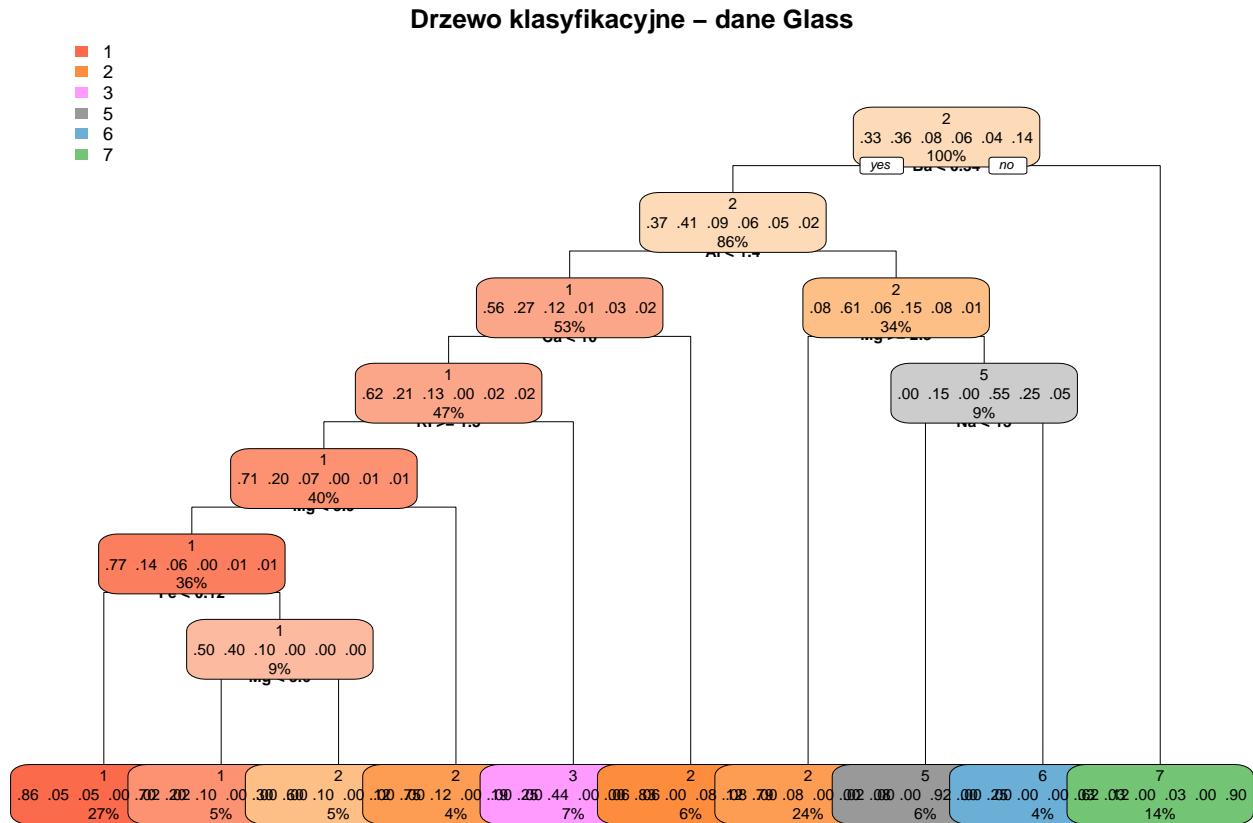
Boxploty zmiennych względem typu szkła



Rysunek 3: Wykresy pułapkowe cech zbioru danych Glass

Wykresy pułapkowe cech zbioru danych Glass (Rysunek 3) przedstawiają rozkład wartości zmiennych chemicznych względem typu szkła. Najlepsze zmienne do rozróżniania klas wydają się być Mg (magnez) i Al (glin). Ich boxploty wykazują różnice w medianach i zakresach między poszczególnymi klasami, co sugeruje, że te pierwiastki mają potencjał dyskryminacyjny. Z

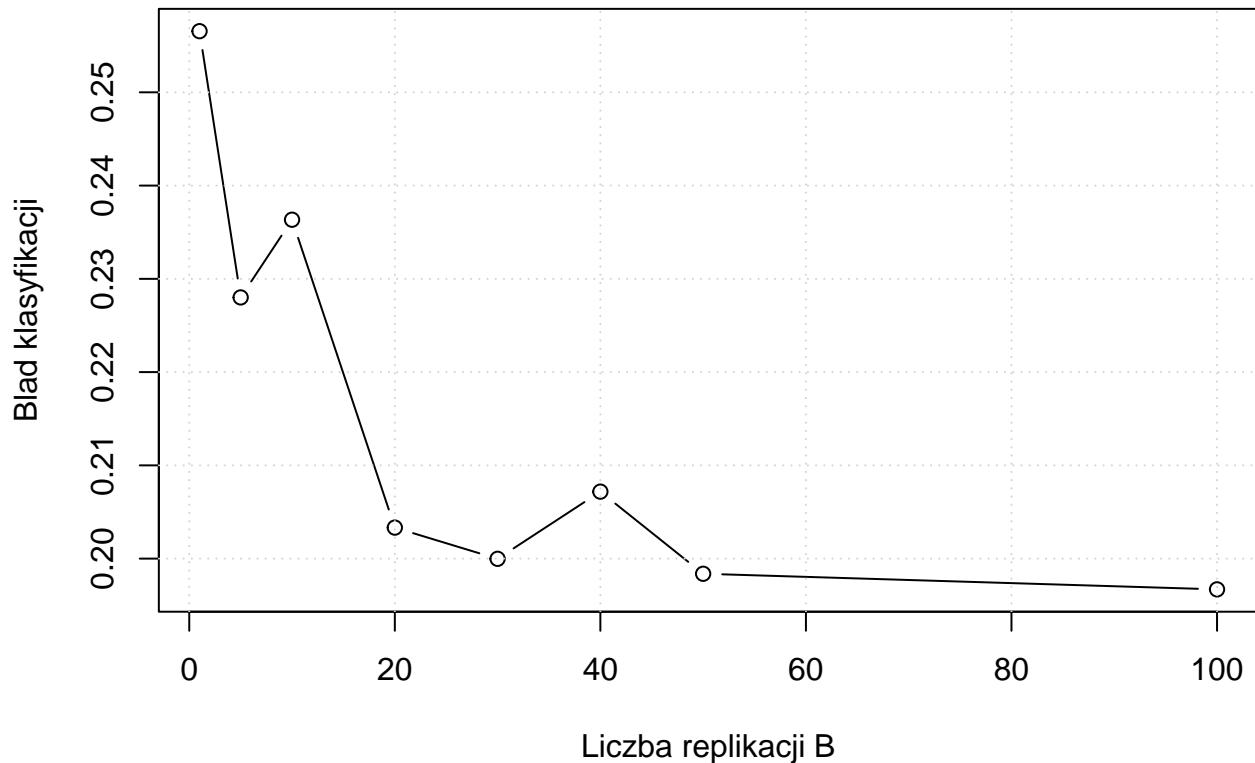
kolej ograniczone zdolności do rozróżniania klas posiada Fe (żelazo) oraz K (potas). Rozkłady wartości są zbliżone dla wszystkich klas, z nakładającymi się boxplotami, co wskazuje na jej niewielki potencjał dyskryminacyjny.



Rysunek 4: Drzewo klasyfikacyjne dla zbioru danych Glass

Rysunek 4 przedstawia strukturę drzewa klasyfikacyjnego zbudowanego dla zbioru danych **Glass**. Drzewo rozpoczyna podział od zmiennej dotyczącej zawartości baru ($Ba < 0.34$), a kolejne rozgałęzienia opierają się na zawartości glinu (Al), magnezu (Mg) oraz wapnia (Ca). Jak pokazano na Rysunku 4, klasa 1 (pomarańczowa) i klasa 2 (żółta) dominują w przewidywaniach, szczególnie w liściach o wysokich wartościach procentowych (np. 27% dla klasy 1 i 24% dla klasy 2). Jednak obecność różnych klas w poszczególnych liściach, takich jak klasy 3, 5, 6 i 7 w mniejszych proporcjach, wskazuje na trudności modelu w jednoznacznym rozróżnianiu niektórych przypadków, co sugeruje pewne nakładanie się cech między klasami.

Bagging – wpływ liczby replikacji



Rysunek 5: Wpływ liczby replikacji na błąd baggingu

Tabela 3: Wpływ liczby replikacji na błąd klasyfikacji

Liczba replikacji B	1	5	10	20	30	40	50	100
Błąd klasyfikacji	25.66%	22.80%	23.63%	20.33%	20.00%	20.72%	19.84%	19.67%

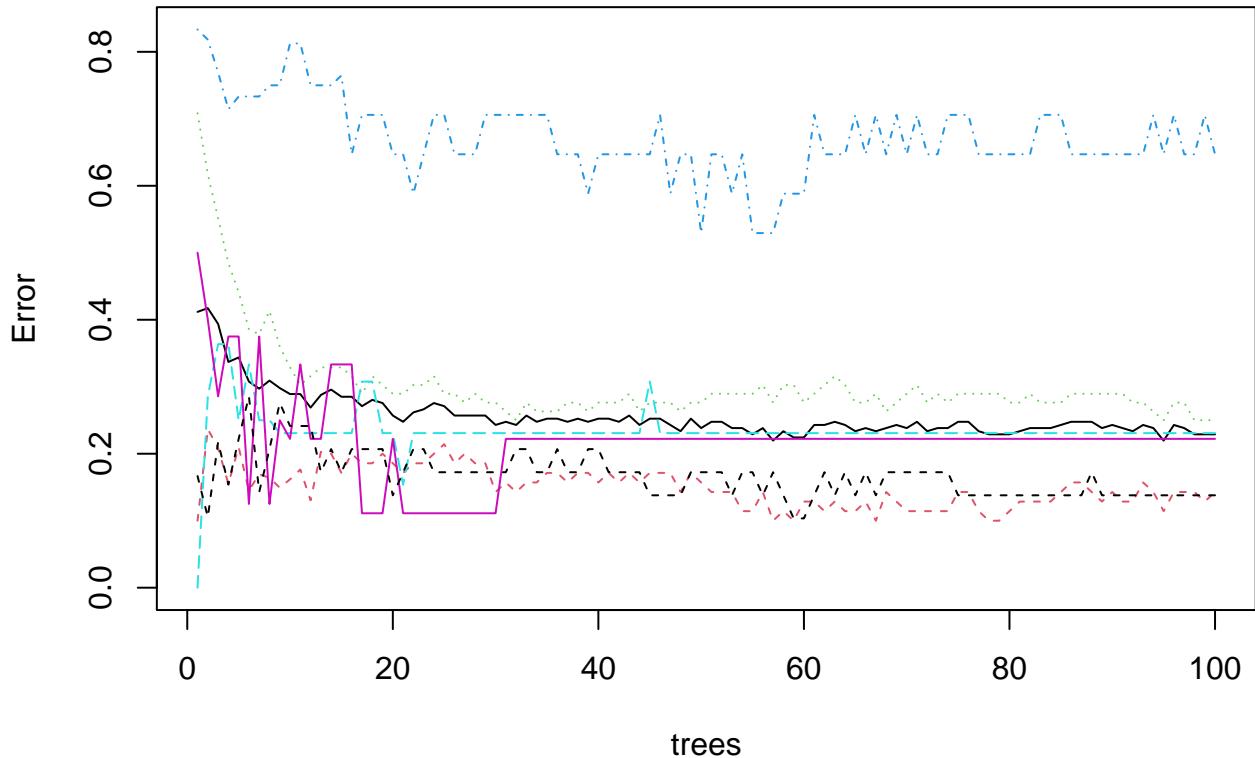
Rysunek 5 i Tabela 3 ilustrują wpływ liczby replikacji (B) na błąd klasyfikacji w metodzie bagging dla zbioru danych. Na Rysunku 5 widać, że błąd klasyfikacji maleje wraz ze wzrostem liczby replikacji B. Tabela 3 potwierdza te trendy: najniższy błąd wynosi 19.67% dla B=100, podczas gdy przy B = 1 błąd wynosi 25.66%.

Tabela 4: Macierz pomyłek - Random Forest dla zbioru danych Glass

		1	2	3	5	6	7
		1	70	0	0	0	0
		2	0	76	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

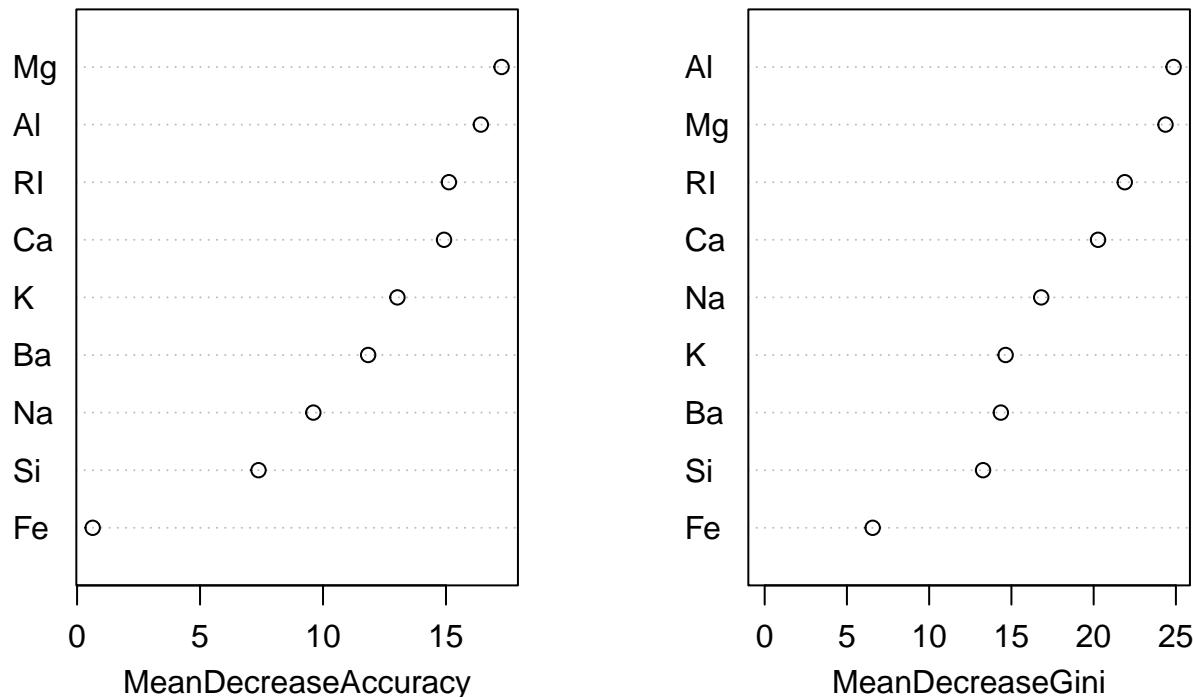
Blad klasyfikacji w Random Forest



Rysunek 6: Błąd modelu Random Forest w zależności od liczby drzew

Wyniki modelu Random Forest dla zbioru danych Glass są przedstawione w Tabeli 4 i Rysunku 6. Tabela 4 (macierz pomyłek) wskazuje, że model doskonale radzi sobie z klasami, ponieważ nie odnotowano żadnych błędów klasyfikacji. Rysunek 6 pokazuje błąd klasyfikacji w zależności od liczby drzew (od 0 do 100), gdzie widać, że początkowo błąd dla poszczególnych jest wysoki i waha się, jednak wraz ze wzrostem liczby drzew, krzywe błędu zbiegają się i stabilizują, co wskazuje na poprawę i stabilizację działania modelu. Zatem model Random Forest wykazuje bardzo wysoką skuteczność w klasyfikacji danych Glass, co jest widoczne zarówno w stabilizacji błędu wraz ze wzrostem liczby drzew, jak i w idealnej macierzy pomyłek.

Ważność zmiennych – Random Forest



Rysunek 7: Ważność zmiennych w modelu Random Forest

Oceniając istotność cech dla modelu Random Forest, Rysunek 7 przedstawia wgląd w ważność zmiennych na podstawie miar `MeanDecreaseAccuracy` i `MeanDecreaseGini`. Z analizy obu wykresów wynika, że zmienne takie jak Mg, RI oraz Al konsekwentnie wykazują największy wpływ na działanie modelu, co świadczy o ich kluczowej roli w klasyfikacji danych Glass. Z kolei zmienna Fe, według obu miar, cechuje się najniższą ważnością.

Tabela 5: Błędy klasyfikacji metod

Metoda	Błąd klasyfikacji (.632+)
Pojedyncze drzewo	30.81%
Bagging	20.08%
Random Forest	15.98%

Tabela 6: Porównanie redukcji błędu

Porównanie	Redukcja błędu (%)
------------	--------------------

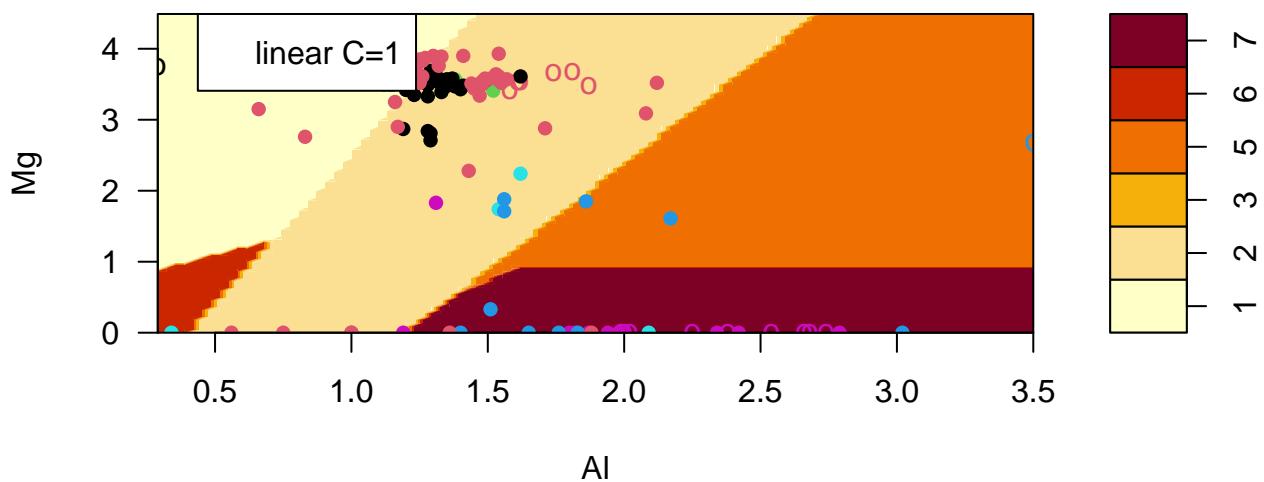
Bagging vs Tree	34.85%
Random Forest vs Tree	48.15%

Badając skuteczność różnych metod klasyfikacji, zauważono wyraźne różnice w ich błędach, co jest przedstawione w Tabeli 5 Pojedyncze drzewo decyzyjne wykazało błąd klasyfikacji na poziomie 30.81%. Znaczącą poprawę odnotowano dla metody Bagging, której błąd wyniósł 20.08%. Najlepsze rezultaty osiągnął model Random Forest, z błędem klasyfikacji na poziomie 15.98%.

Analizując redukcję błędu względem pojedynczego drzewa, co szczegółowo przedstawia Tabela 6, Bagging zmniejszył błąd o 34.85%. Random Forest okazał się jeszcze skuteczniejszy, redukując błąd aż o 48.15%. Zatem metody zespołowe, a w szczególności Random Forest, istotnie przewyższają pojedyncze drzewa w klasyfikacji danych **Glass**, oferując znacznie wyższą precyzję.

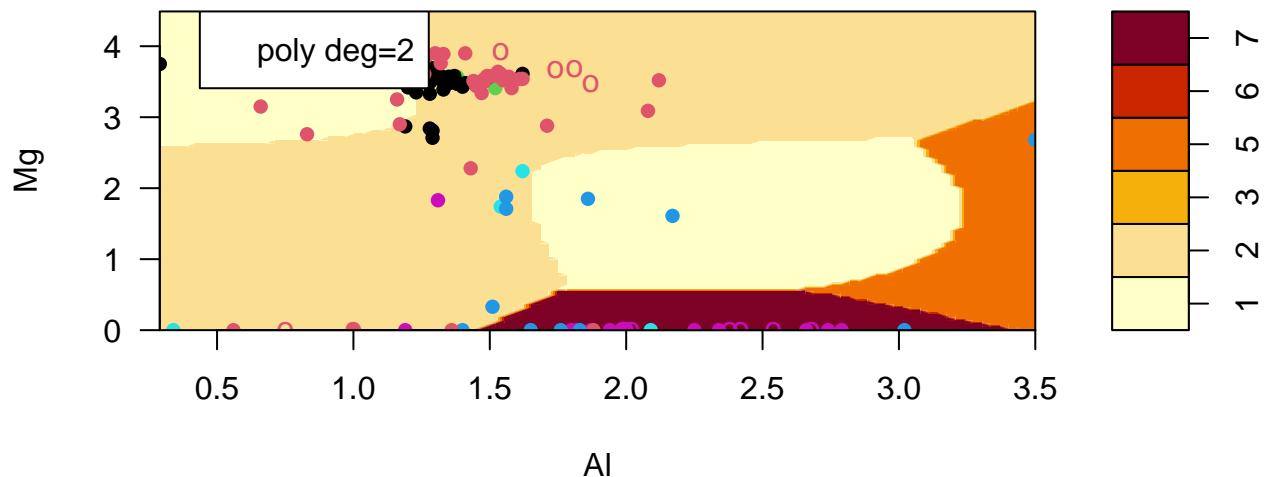
1.3 Metoda wektorów nośnych (SVM)

SVM classification plot



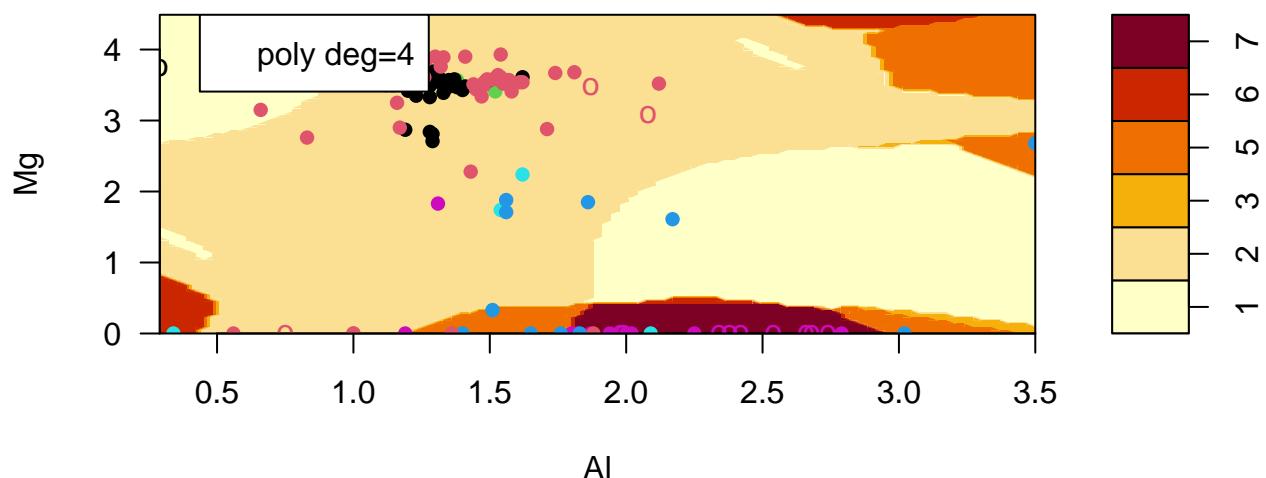
Rysunek 8: Wizualizacja klasyfikatorów SVM dla różnych funkcji jądrowych i parametrów

SVM classification plot



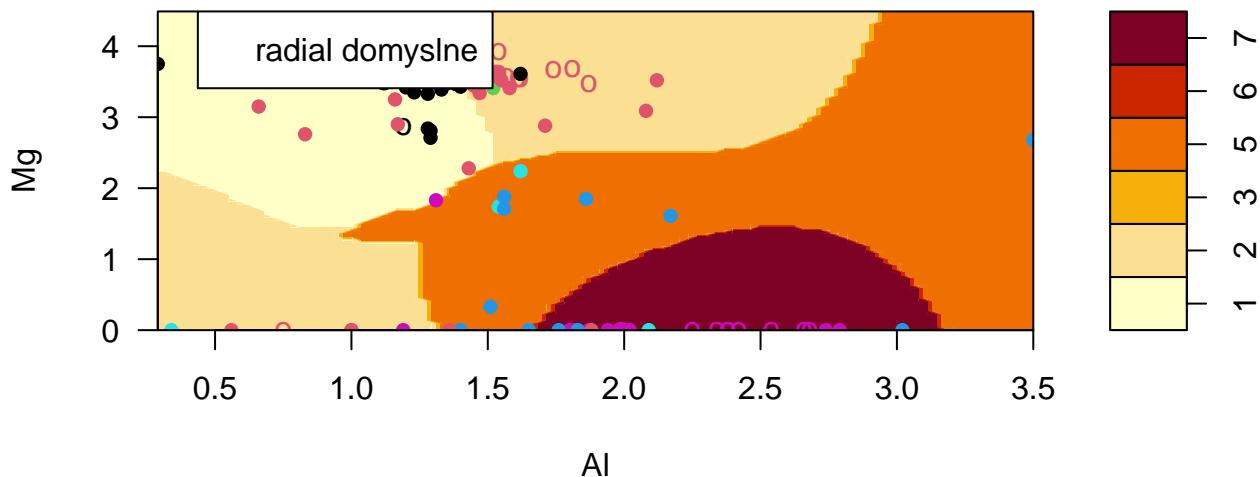
Rysunek 9: Wizualizacja klasyfikatorów SVM dla różnych funkcji jądrowych i parametrów

SVM classification plot



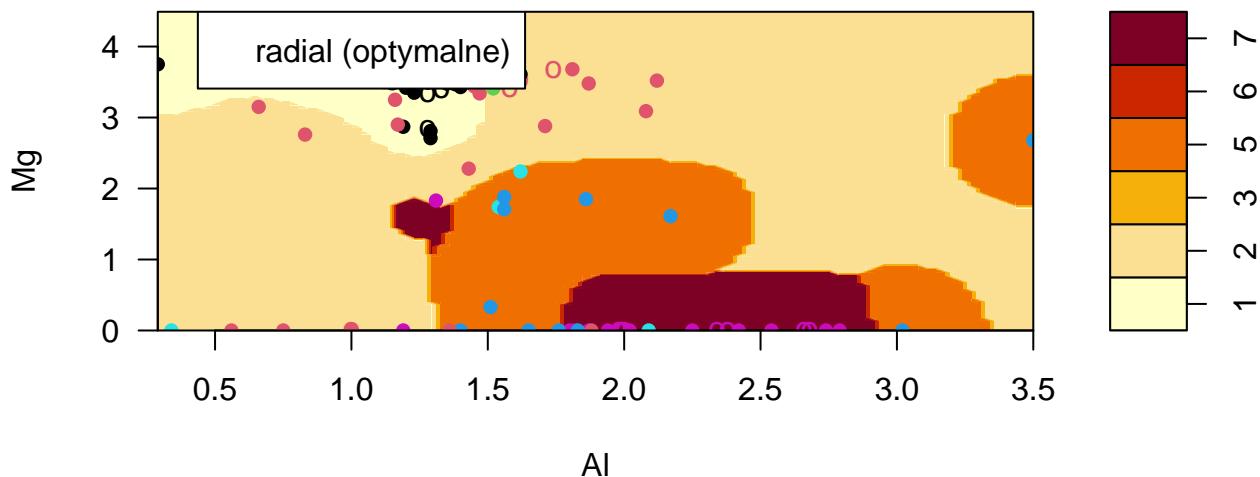
Rysunek 10: Wizualizacja klasyfikatorów SVM dla różnych funkcji jądrowych i parametrów

SVM classification plot



Rysunek 11: Wizualizacja klasyfikatorów SVM dla różnych funkcji jądrowych i parametrów

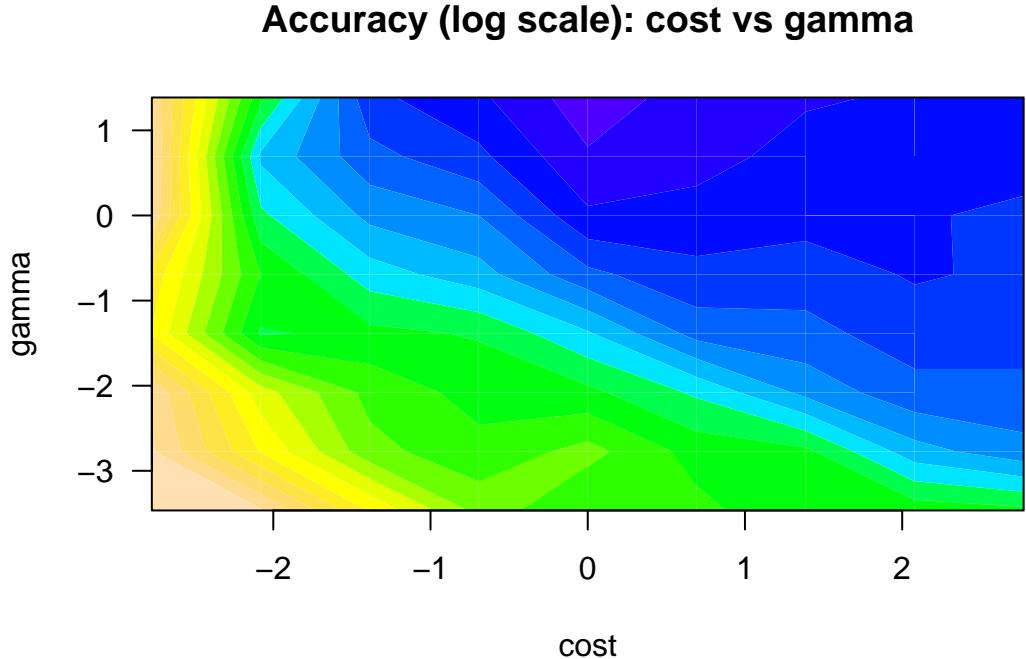
SVM classification plot



Rysunek 12: Wizualizacja klasyfikatorów SVM dla różnych funkcji jądrowych i parametrów

Rysunki od 8 do 12 przedstawiają wizualizacje działania klasyfikatorów SVM z różnymi funkcjami jądrowymi i parametrami, bazującymi na zmiennych Mg i Al ze zbioru danych Glass. Na każdym z wykresów widoczne są punkty danych (próbki szkła) oraz obszary decyzyjne, reprezentujące przewidywane klasy. Obserwujemy, że jądro liniowe tworzy proste granice, podczas gdy jądra wielomianowe generują coraz bardziej złożone, zakrzywione obszary. Jądra radialne pozwalają na tworzenie najbardziej elastycznych i skomplikowanych granic, co jest szczególnie widoczne w przypadku jądra radialnego z optymalnymi parametrami, gdzie granice decyzyjne są precyzyjnie dopasowane do grup punktów. Ogólnie rzecz biorąc, im

bardziej złożone jadro, tym lepsze dopasowanie do nieliniowej struktury danych i potencjalnie lepsza separacja klas.



Rysunek 13: Wizualizacja dokładności klasyfikacji w przestrzeni parametrów (cost, gamma)

Tabela 7: Najlepsze parametry radialnego SVM i odpowiadająca im dokładność

Parametr	Wartość
Najlepszy cost (C)	1
Najlepszy gamma	4

Rysunek 13 wizualizuje dokładność klasyfikacji w przestrzeni parametrów `cost` i `gamma` dla modelu SVM, możemy zaobserwować, jak te parametry wpływają na skuteczność klasyfikatora. Z wykresu wynika, że najwyższa dokładność koncentruje się niedaleko górnego lewego rogu, gdzie wartości `cost` są niższe i wartości `gamma` są wyższe. To sugeruje, że dla tego zbioru danych model SVM osiąga najlepsze wyniki przy niższych wartościach parametru kary C (`cost`) i wyższych wartościach `gamma`. Tabela 7 wskazuje, że najlepsze parametry to `cost` (C) równy 1 oraz `gamma` równy 4. Ta kombinacja parametrów zapewnia optymalną dokładność modelu radialnego SVM dla zbioru danych Glass.

Tabela 8: Porównanie skuteczności różnych funkcji jądro-wych i parametrów C

Model	Dokładność
SVM liniowy (C=0.1)	48.61%
SVM liniowy (C=1)	51.39%
SVM liniowy (C=10)	51.39%
SVM poly deg=2	50.00%
SVM poly deg=4	47.22%
SVM radial (domyślne)	52.78%
SVM radial (optymalne)	54.17%

Tabela 9: Macierz pomyłek: SVM liniowy (C = 0.1, 1, 10)

	1	2	3	5	6	7
1	9	2	3	0	0	0
2	14	22	5	0	4	5
3	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	1	0	2	1	4

Tabela 10: Macierz pomyłek: SVM poly (deg=2)

	1	2	3	5	6	7
1	11	3	3	0	0	2
2	12	21	5	0	5	3
3	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	1	0	2	0	4

Tabela 11: Macierz pomyłek: SVM poly (deg=4)

	1	2	3	5	6	7
1	4	0	2	0	0	0
2	19	24	6	0	5	5
3	0	0	0	0	0	0
5	0	0	0	2	0	0
6	0	0	0	0	0	0

$$\begin{array}{ccccccc} 7 & 0 & 1 & 0 & 0 & 0 & 4 \\ \hline \end{array}$$

Tabela 12: Macierz pomyłek: SVM radial (domyślne)

	1	2	3	5	6	7
1	17	5	5	0	1	1
2	6	16	3	0	2	2
3	0	0	0	0	0	0
5	0	3	0	1	2	2
6	0	0	0	0	0	0
7	0	1	0	1	0	4

Tabela 13: Macierz pomyłek: SVM radial (optymalne)

	1	2	3	5	6	7
1	19	7	5	0	0	1
2	4	14	3	0	3	2
3	0	0	0	0	0	0
5	0	3	0	2	2	2
6	0	0	0	0	0	0
7	0	1	0	0	0	4

Przeglądając się skuteczności różnych funkcji jądrowych i parametrów C, widać wyraźne zróżnicowanie ich dokładności, co jest szczegółowo przedstawione w Tabeli 8. Wyniki wskazują, że liniowe SVM (dla $C=0.1$, $C=1$, $C=10$) osiągnęły taką samą dokładność w zakresie 48.61–51.39%. Modele z jądrem wielomianowym (stopnia 2 i 4) uzyskały dokładności odpowiednio 50% i 47.22%. Najlepsze wyniki odnotowano dla jądra radialnego, gdzie domyślny model osiągnął 52.78% dokładności, a model zoptymalizowany 54.17%, widoczna jest więc poprawa. Tabele 9, 10, 11, 12 i 13, które prezentują macierze pomyłek dla różnych konfiguracji SVM, można zauważyć, jak zmienia się zdolność modelu do poprawnej klasyfikacji poszczególnych klas. Na przykład, dla SVM liniowego klasa 1 jest mylona z klasą 2, a klasa 2 z klasą 1. Jądra wielomianowe wykazują nieco lepsze rezultaty w rozróżnianiu klas niż liniowe. Natomiast dla SVM radialnego pomyłki między klasami są najmniejsze, choć nadal występują.

Wybór funkcji jądrowej oraz wybór parametru kosztu C w istotnym stopniu wpływają na dokładność metody SVM. Jak widać w Tabeli 6, zmiana parametru kosztu C z 0.1 na 1 lub 10 nie przynosi dużej poprawy dokładności, utrzymując ją na podobnym poziomie. Natomiast zmiana funkcji jądrowej z liniowej na wielomianową nieco pogarsza dokładność. Co więcej, w przypadku jądra wielomianowego, stopień 2 spisał się lepiej niż stopień 4. Natomiast przejście do jądra radialnego przynosi największą poprawę dodatkowo optymalizacja parametrów przyniosła korzyść i pozwoliła na poprawę skuteczności skonstruowanego klasyfikatora.

1.4 Porównanie skuteczności metod

Porównując wyniki uzyskane w analizie metod ensemble learning i klasyfikatorów SVM, można zauważać wyraźne różnice w skuteczności działania poszczególnych podejść.

Zdecydowanie najlepsze rezultaty osiągnęła metoda Random Forest, która przewyższyła zarówno klasyczne drzewo decyzyjne, jak i bagging. Random Forest okazał się nie tylko najbardziej precyzyjny, ale również najbardziej stabilny — model ten skutecznie rozpoznawał wszystkie klasy i dobrze radził sobie z różnorodnością danych. W odróżnieniu od pojedynczego drzewa, które miało trudności z uogólnieniem danych.

Z kolei metody oparte na SVM, mimo zastosowania różnych funkcji jądrowych (liniowe, wielomianowe i radialne) oraz optymalizacji parametrów, okazały się mniej skuteczne. Choć najbardziej zaawansowany wariant – SVM z radialnym jądrem i dobranymi parametrami – poradził sobie lepiej niż pozostałe konfiguracje SVM, nadal nie osiągnął poziomu skuteczności porównywalnego z Random Forest. Klasyfikatory SVM miały trudności z dokładnym rozróżnianiem niektórych klas, zwłaszcza tych mniej licznych, co negatywnie wpłynęło na ogólną trafność modelu.

Można więc stwierdzić, że Random Forest zdecydowanie najlepiej poradził sobie z klasyfikacją danych Glass, wyraźnie przewyższając zarówno bagging, jak i wszystkie warianty SVM dla zbioru danych ‘Glass’.

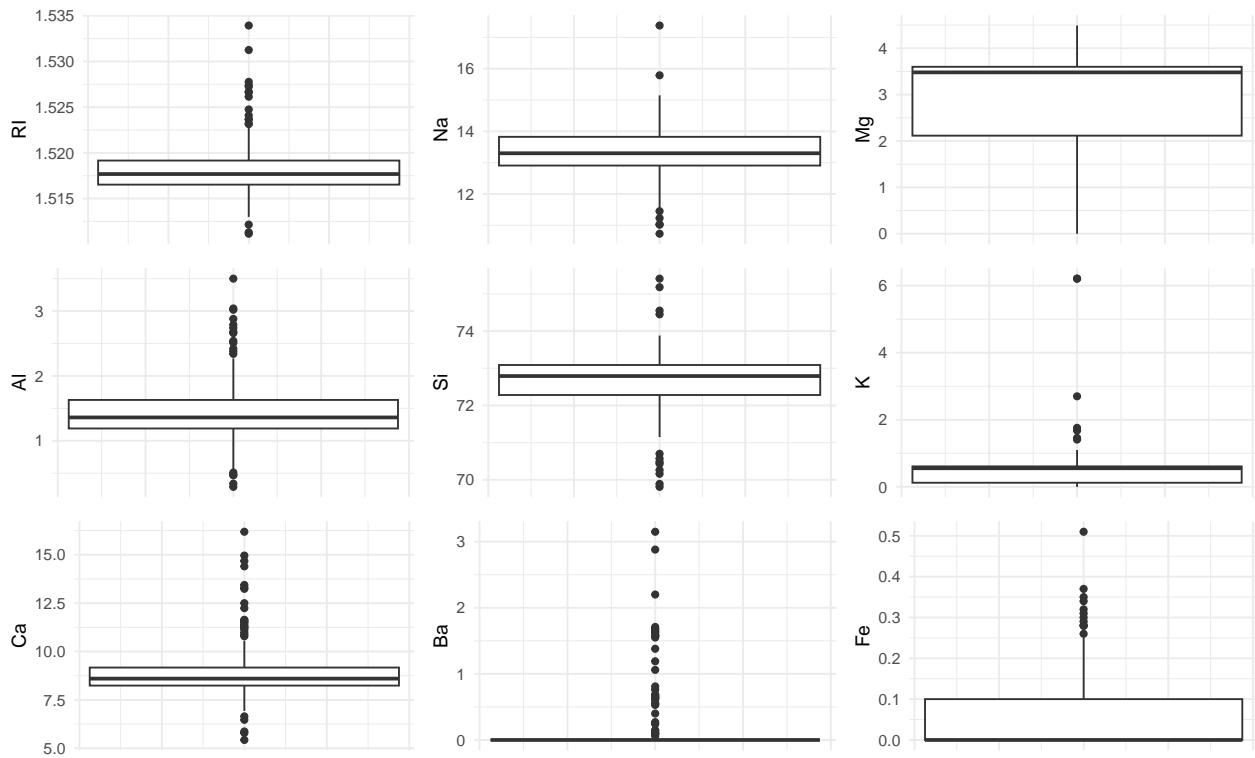
2 Analiza skupień - algorytmy grupujące i hierarchiczne

2.1 Przygotowanie danych

W tym zadaniu skupimy się na analizie skupień. Do jej wykonania wykorzystamy ten sam zbiór co w zadaniu 2 z listy 3, czyli zbiór Glass (mlbench), który opisuje dane identyfikacyjne szkła używane oraz potrzebne przy śledztwach kryminalistycznych.

Zauważmy, że wszystkie dane w naszym zbiorze poza zmienną reprezentującą klasy - Type są numeryczne. Zatem usunięcie z analizy zmiennej grupującej zawierającej etykiety klas zrobimy poprzez przeprowadzenie analizy na zbiorze zmiennych numerycznych.

Pomijamy szczegółowy opis danych, ponieważ pracujemy na dobrze znanym i wcześniej analizowanym zbiorze (z zadania 2 listy 3 oraz zadania 1 z listy 4). Kluczowe jest jednak powtórne sprawdzenie potrzeby standaryzacji danych, gdyż ten krok ma fundamentalne znaczenie dla wiarygodności naszej analizy skupień.



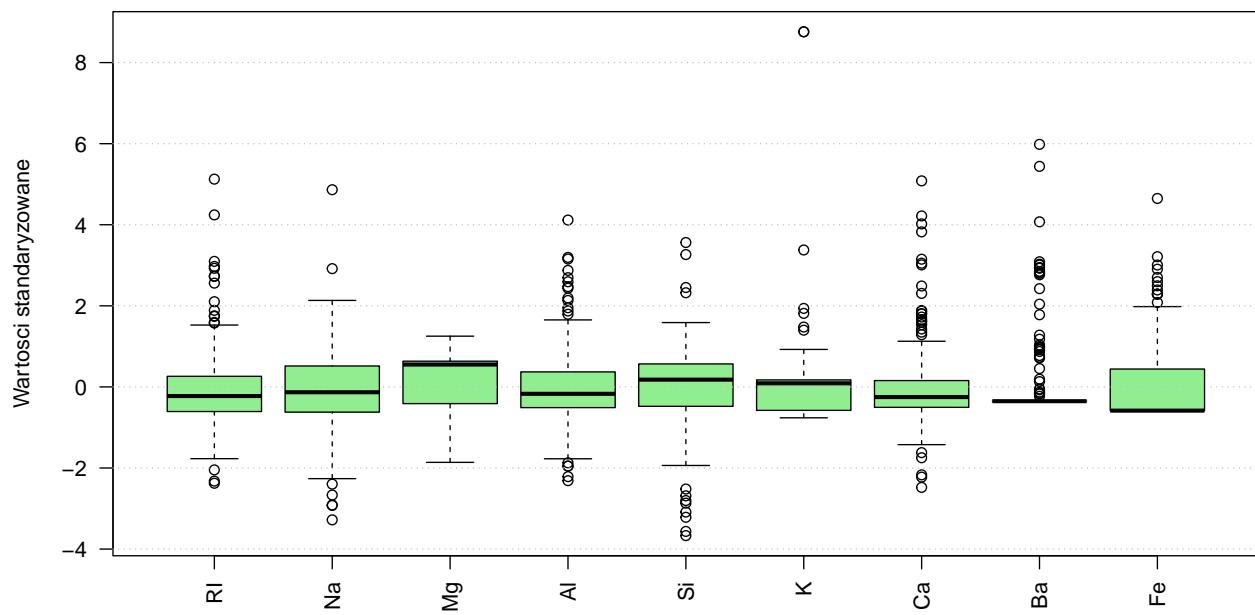
Rysunek 14: Wykresy pudełkowe względem zmiennych

Tabela 14: Wariancje poszczególnych zmiennych numerycznych

Zmienna	Wariancja
RI	0.0000092
Na	0.6668414
Mg	2.0805404
Al	0.2492702
Si	0.5999212
K	0.4253542
Ca	2.0253658
Ba	0.2472270
Fe	0.0094943

Biorąc pod uwagę wartości przedstawione w Tabeli 14, która obrazuje wariancje analizowanych zmiennych, oraz Rysunku 14, prezentującego wykresy pudełkowe tych zmiennych, standaryzacja danych jest konieczna. Widzimy, że wartości wahają się od zaledwie 0.0000092 dla RI do ponad 2 dla Mg oraz Ca, co jasno wskazuje na potrzebę standaryzacji (wizualizacja po standaryzacji - Rysunek 15).

Wykresy pudełkowe po standaryzacji



Rysunek 15: Wykresy pudełkowe względem zmiennych

2.2 Wizualizacja wyników grupowania

Skupimy się na dwóch głównych algorytmach grupowania: K-średnich (K-means) oraz hierarchicznego. Przyjmijmy liczbę skupień K jako równą rzeczywistej liczbie klas.

Rzeczywista liczba klas (K) w zbiorze Glass: 6

Wizualizacja skupień (K-Means, K=6) i rzeczywistych klas (PCA)

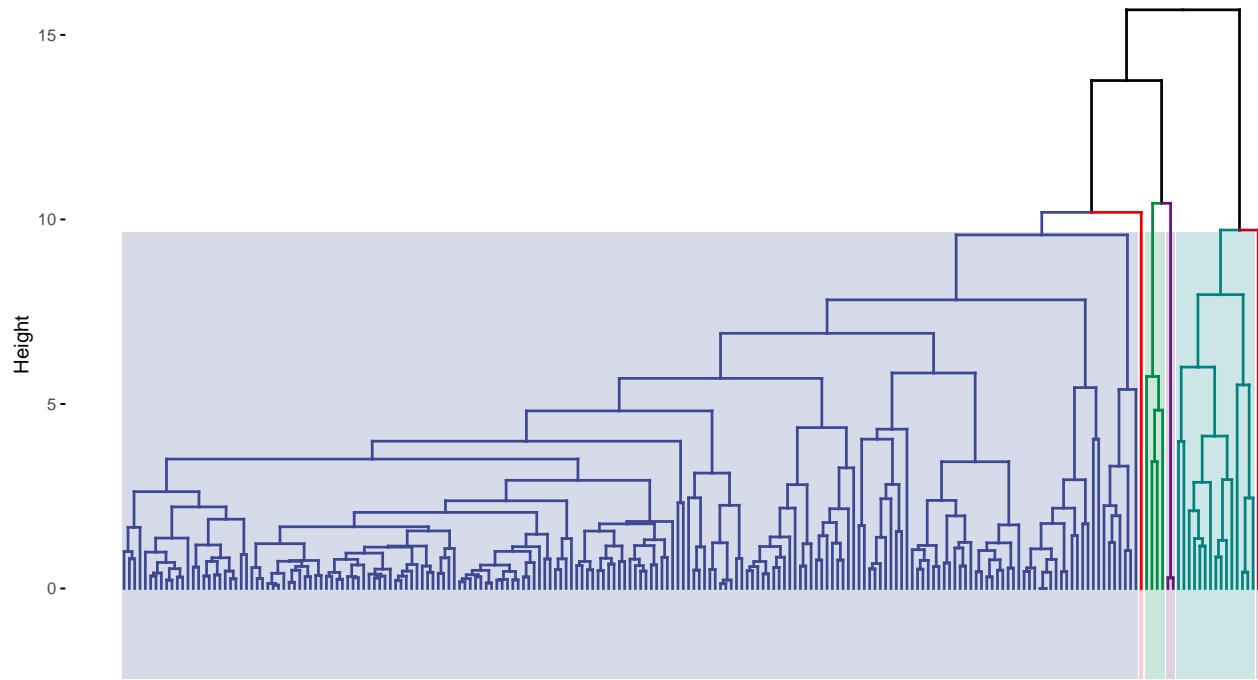
Kolor: Przypisane skupienia, Kształt: Rzeczywiste klasy



Rysunek 16: Wykres rozrzutu zmiennych po grupowaniu algorytmem K-means

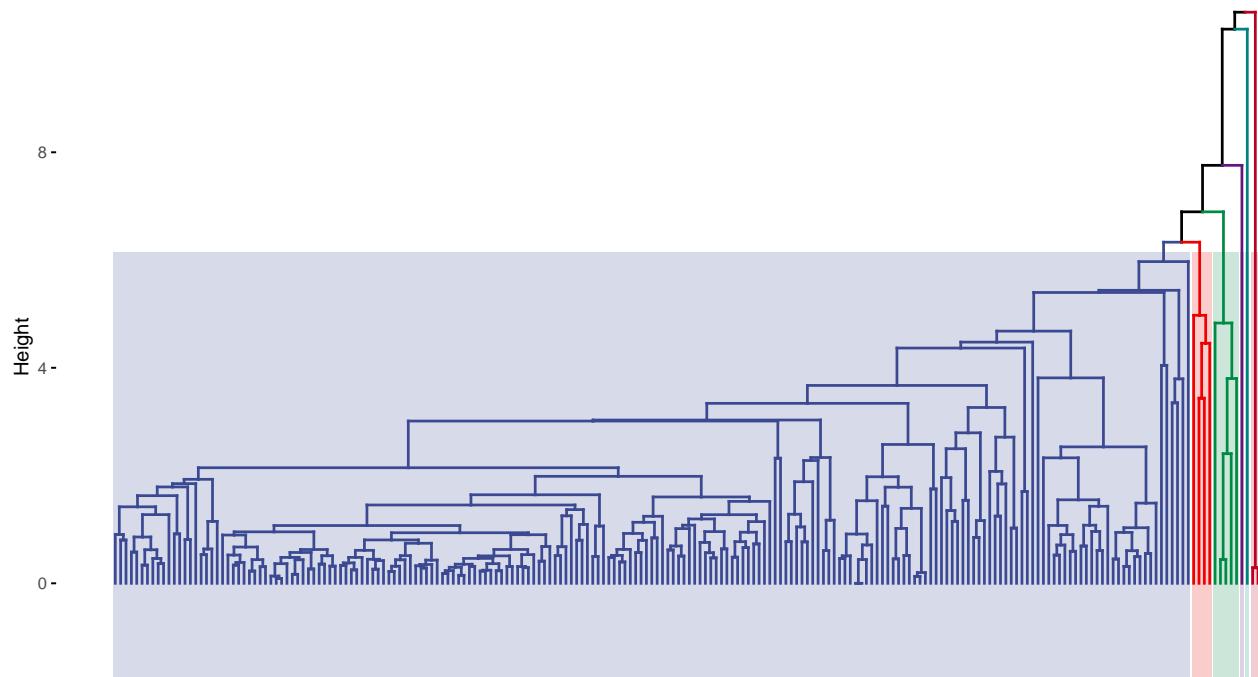
Na Rysunku 16 widać, że redukcja wymiarów przy pomocy PCA pozwoliła zobrazować strukturę danych w dwóch głównych komponentach. Widoczna jest częściowa zgodność między skupieniami a klasami, jednak pewne klasy są rozproszone po różnych klastrach, co może świadczyć o ich słabszej separowalności w przestrzeni cech.

Dendrogram (Agnes, Complete) – 6 skupien



Rysunek 17: Dendrogram Agnes - Complete

Dendrogram (Agnes, Average) – 6 skupien



Rysunek 18: Dendrogram Agnes - Average

Dendrogram (Agnes, Single) – 6 skupień

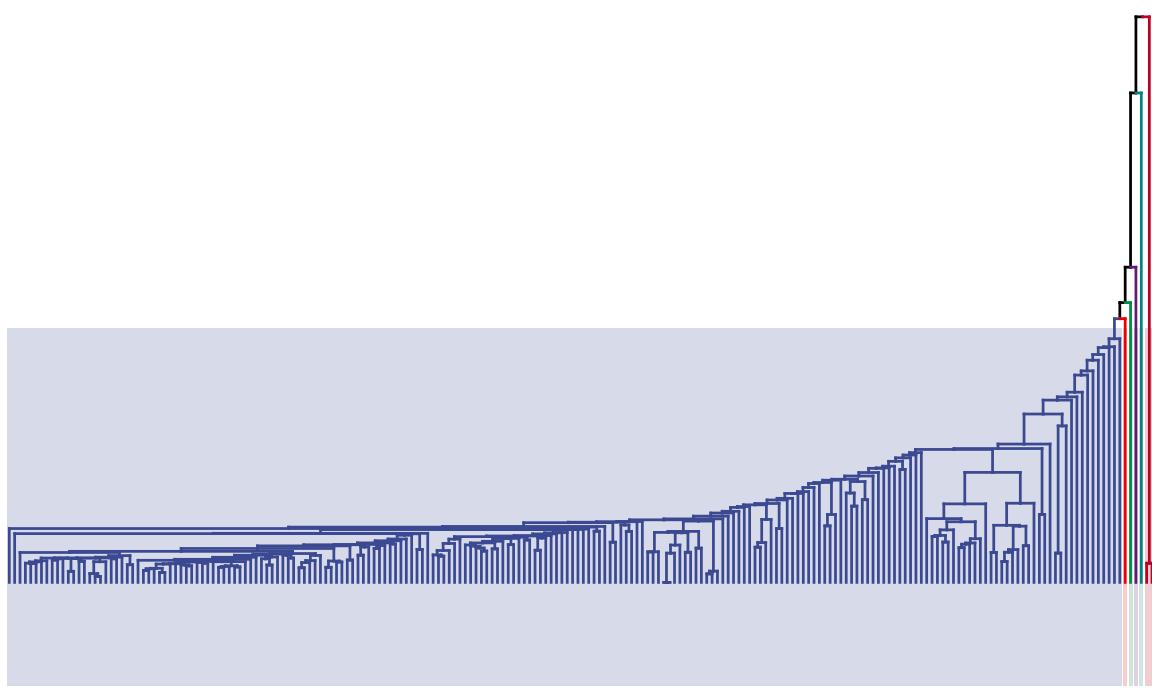
9 -

6 -

3 -

0 -

Height

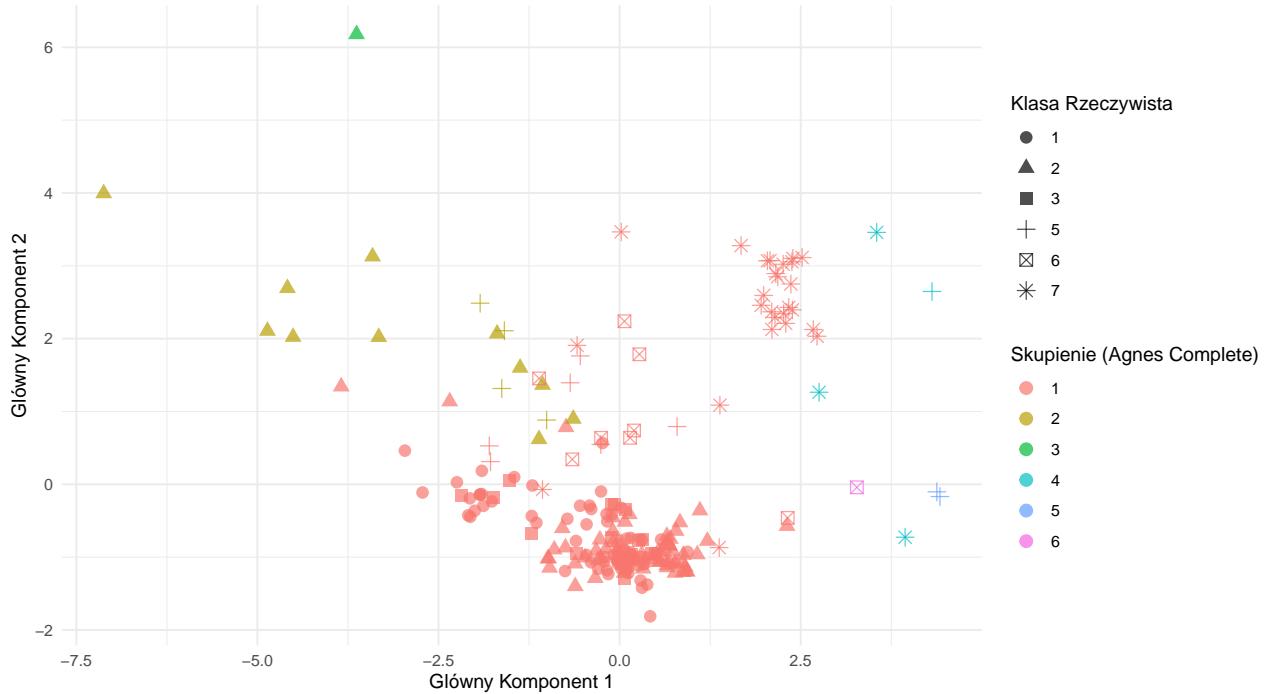


Rysunek 19: Dendrogram Agnes - Single

Dendrogramy prezentują hierarchiczne grupowanie danych przy użyciu różnych metod łączenia: complete (Rysunek 17), average (Rysunek 18) oraz single (Rysunek 19) linkage. Wszystkie zostały „ucięte” na poziomie 6 skupień, zgodnie z rzeczywistą liczbą klas. Metoda complete wykazuje bardziej wyraźną i zbalansowaną strukturę skupień, natomiast single linkage prowadzi do nieco bardziej niestabilnego i „łańcuchowego” łączenia obiektów, co może skutkować mniej czytelnym podziałem.

Skupienia (Agnes Complete, K=6) vs. Rzeczywiste Klasy (PCA)

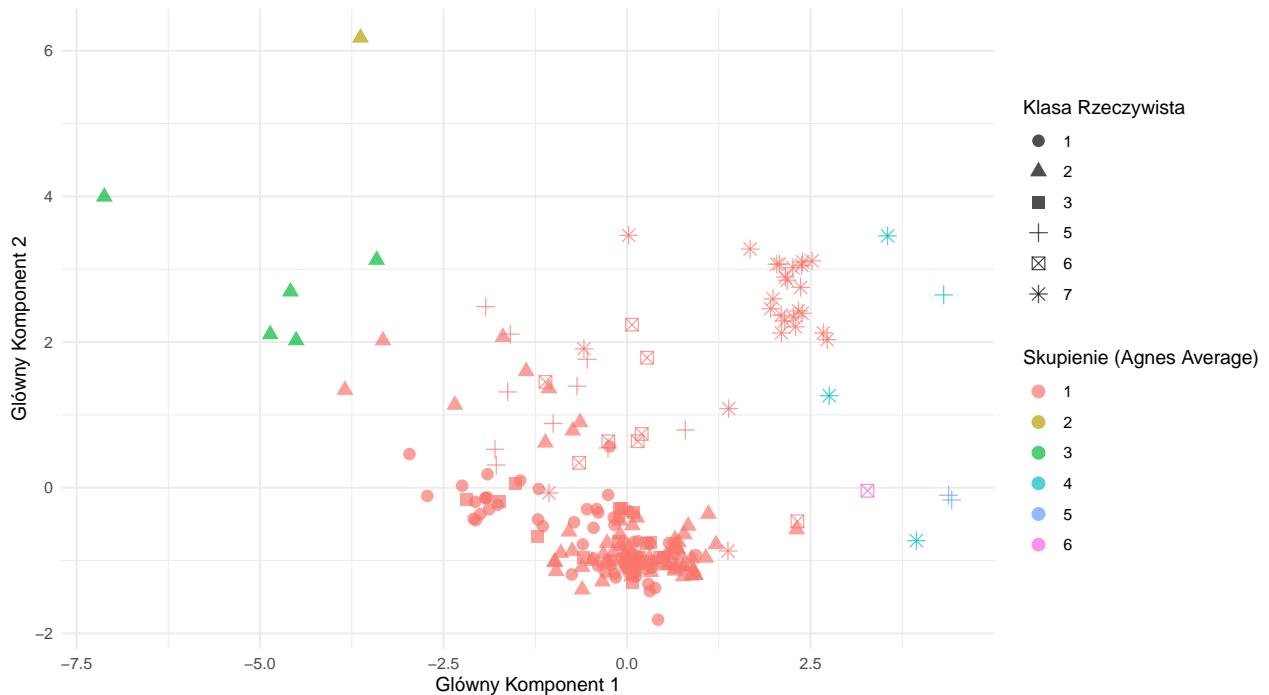
Kolor: Przypisane skupienia (Agnes), Kształt: Rzeczywiste klasy



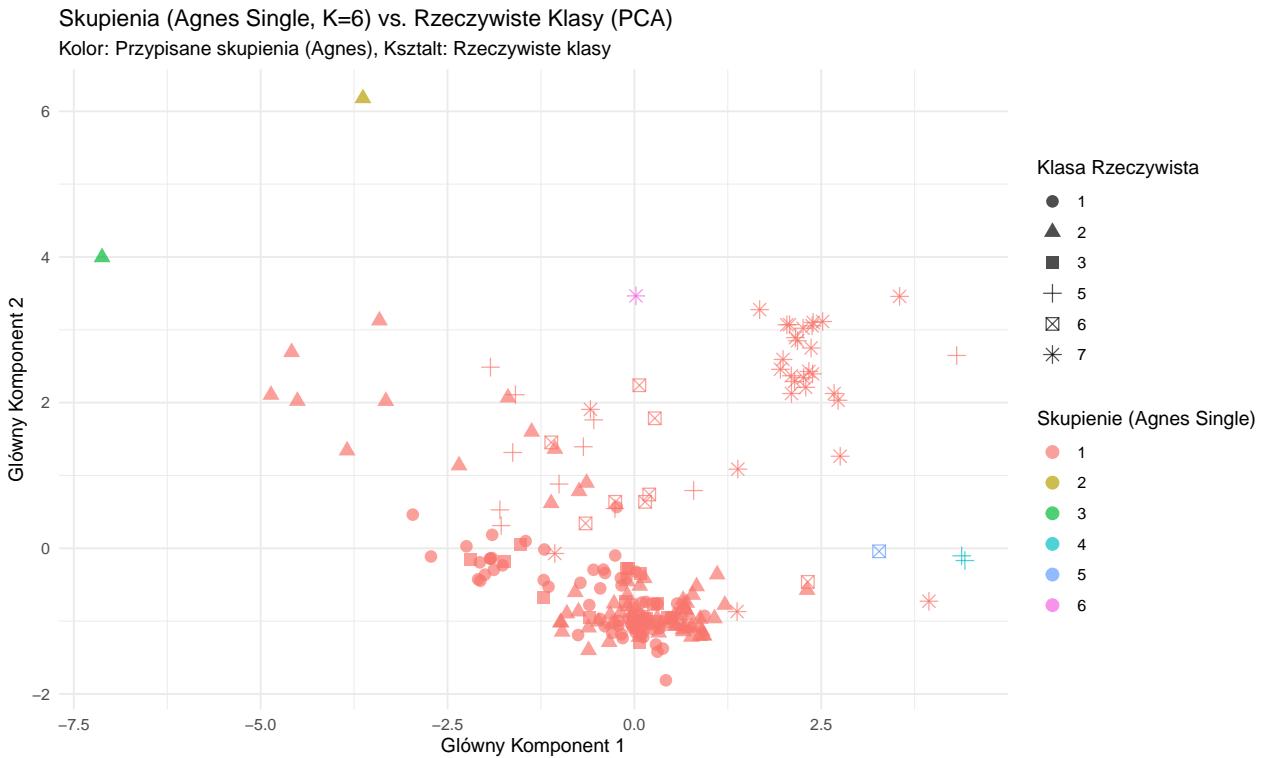
Rysunek 20: Wykres rozrzutu Agnes Complete poprzez PCA

Skupienia (Agnes Average, K=6) vs. Rzeczywiste Klasy (PCA)

Kolor: Przypisane skupienia (Agnes), Kształt: Rzeczywiste klasy



Rysunek 21: Wykres rozrzutu Agnes Average z PCA



Rysunek 22: Wykres rozrzutu Agnes Single poprzez PCA

Każdy z wykresów odpowiada innej metodzie łączenia: complete (Rysunek 20), average (Rysunek 21) i single (Rysunek 22). Kolory wskazują przypisane skupienia, a kształty – rzeczywiste klasy. Metoda complete osiąga stosunkowo najlepsze dopasowanie skupień do klas, natomiast average i szczególnie single pokazują większe rozproszenie obiektów tej samej klasy między różnymi skupieniami. Może to sugerować, że complete linkage lepiej radzi sobie z zachowaniem struktury klas w tym zbiorze.

Analiza wyników grupowania dla zbioru danych Glass pozwala ocenić podstawowe własności uzyskanych skupień, takie jak ich zwartość, jednorodność i separacja. W przypadku algorytmu K-średnich, skupienia są względnie zwarte i częściowo odzwierciedlają rzeczywistą strukturę klas. Jednakże, na wykresie rozrzutu po redukcji wymiarów metodą PCA widoczna jest jedynie umiarkowana separacja. Fakt, że niektóre klasy są rozproszone po różnych klastrach, może wskazywać na ich mniejsze zróżnicowanie w przestrzeni cech.

W analizie hierarchicznej zaś zauważalne są istotne różnice wynikające z zastosowanych metod łączenia. Dendrogramy jednoznacznie pokazują, że metoda complete linkage prowadzi do najlepiej zorganizowanej struktury hierarchicznej, w której skupienia są względnie dobrze wyodrębnione i równomiernie rozłożone. Metoda average linkage daje efekty umiarkowane, stanowiąc pewien kompromis, natomiast single linkage tworzy niestabilne, rozcięgnięte klastry, co świadczy o niskiej zwartości i słabej separacji, wynikającej z tak zwanego efektu chainingu.

Wizualizacje skupień hierarchicznych na płaszczyźnie PCA dodatkowo potwierdzają te obserwacje: najlepszą zgodność między przypisanymi skupieniami a rzeczywistymi klasami można

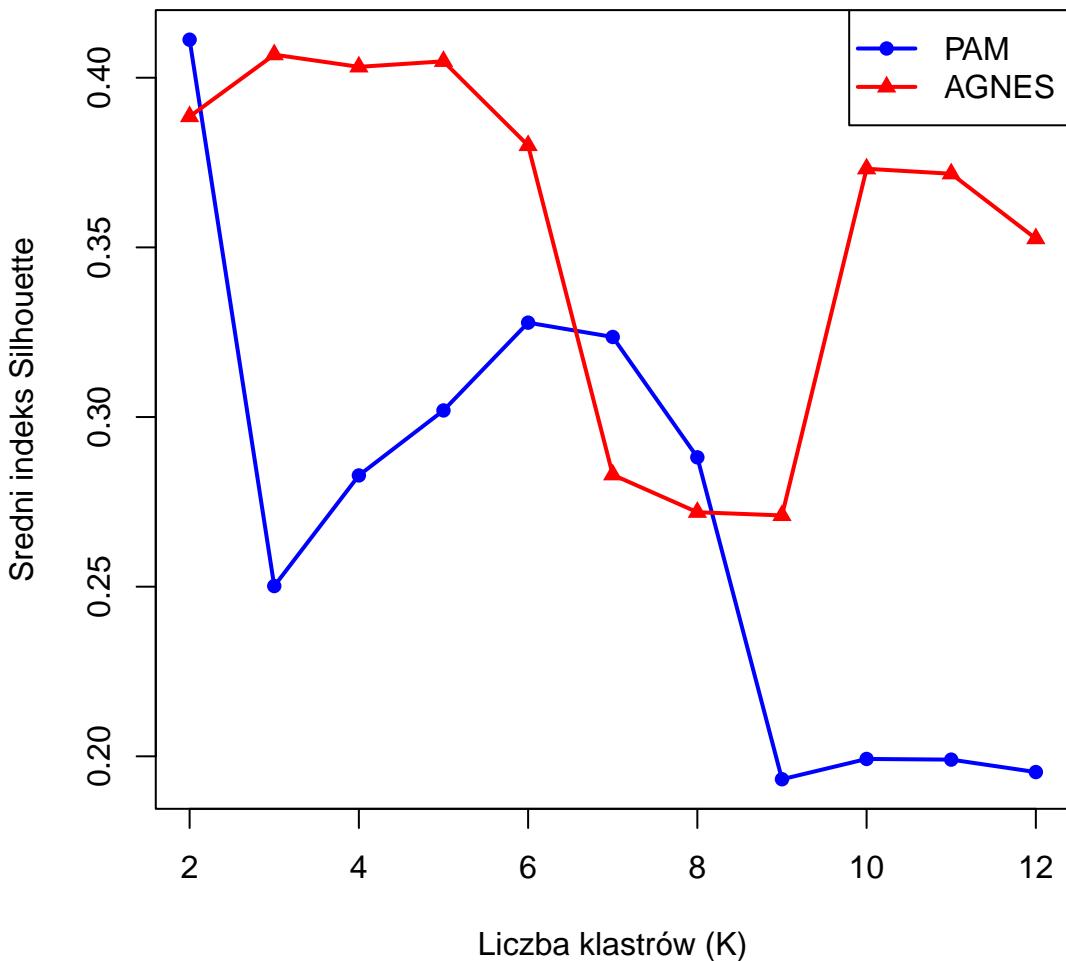
zaobserwować dla metody complete linkage, gdzie obiekty jednej klasy często trafiają do tego samego klastra. Metoda average linkage wykazuje nieco niższą zgodność, a w przypadku single linkage obiekty tej samej klasy są silnie rozproszone, co przekłada się na niską zgodność z rzeczywistą strukturą danych.

Podsumowując, otrzymany podział na skupienia zgadza się z rzeczywistą przynależnością obiektów do klas jedynie częściowo. Najlepsze wyniki uzyskano przy zastosowaniu metody complete linkage oraz algorytmu K-średnich, choć nawet w tych przypadkach występują wyraźne rozbieżności, co sugeruje, że klasy w zbiorze Glass nie są idealnie separowalne w analizowanej przestrzeni cech.

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

Weźmy zakres dla K od 2 do 12, w celu określenia który z algorytmów PAM oraz AGNES lepiej poradził sobie z grupowaniem ustandaryzowanych danych Glass oraz jakie K jest najbardziej optymalne. Weźmiemy AGNES complete, ponieważ jak pokazała powyższa analiza jest to najlepszy możliwy rodzaj algorytmu AGNES.

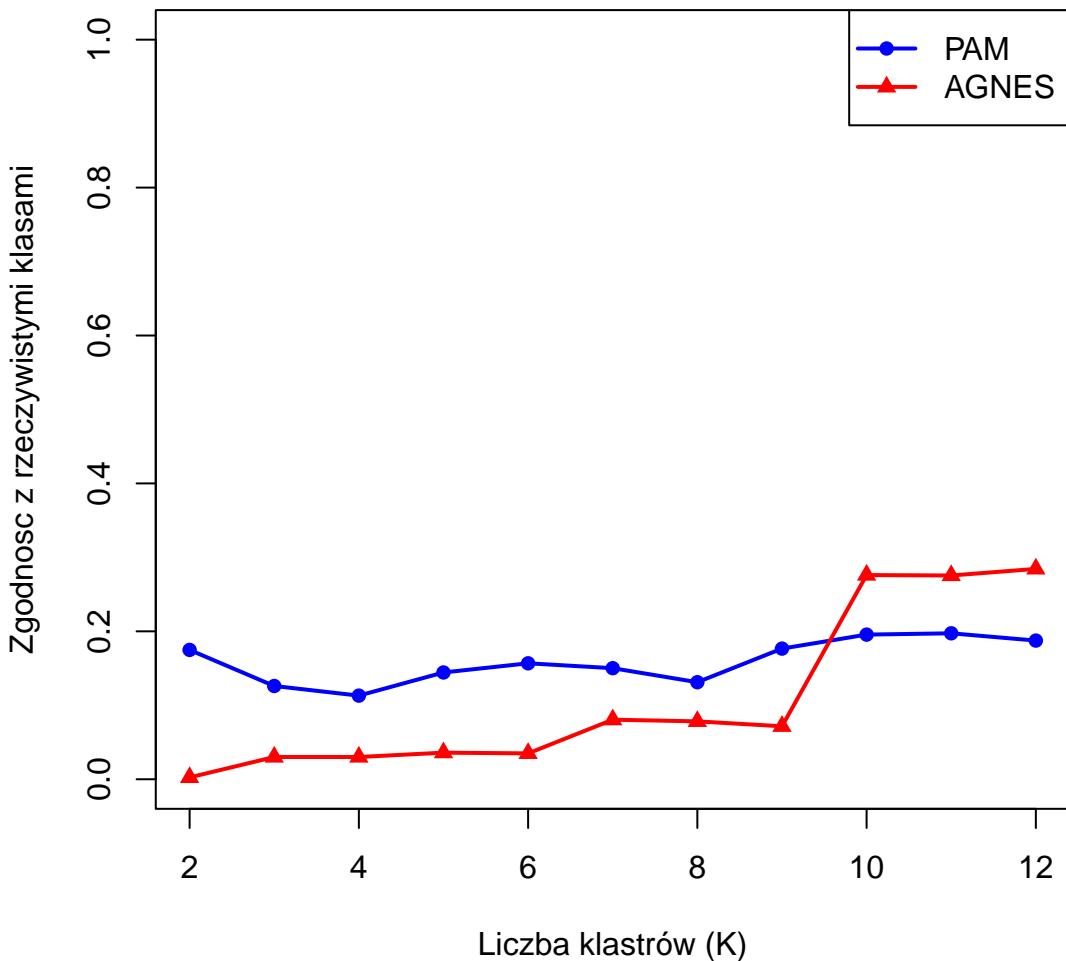
Porównanie silhouette: PAM vs AGNES



Rysunek 23: Wykres porównania Silhouette - PAM vs AGNES

Wykres 23 ilustruje zachowanie wskaźnika wewnętrznego – średniego indeksu Silhouette. Dla algorytmu AGNES (linia czerwona), obserwujemy początkowy wzrost wartości indeksu, z osiągnięciem maksimum w przedziale K od 3 do 5, gdzie wartość ta kształtuje się na poziomie około 0.40-0.41. Następnie następuje spadek, a potem ponowny wzrost dla K od 10 do 11 do około 0.37. W przypadku algorytmu PAM (linia niebieska), najwyższą wartość około 0.41 odnotowujemy dla K=2, po czym następuje gwałtowny spadek do około 0.25 dla K=3, a następnie stopniowy wzrost do około 0.32 dla K=6 i K=7. Wartości indeksu dla PAM drastycznie maleją dla K powyżej 8. Na podstawie tych danych, AGNES wydaje się być bardziej stabilny i osiąga wyższe wartości Silhouette w większości analizowanego zakresu K, co sugeruje tworzenie bardziej spójnych i lepiej oddzielonych skupień wewnętrznie. Optymalne K dla AGNES, z perspektywy Silhouette, wydaje się leżeć w przedziale od 4 do 5.

Porównanie zgodności z klasami: PAM vs AGNES



Rysunek 24: Wykres porównania zgodności z rzeczywistymi klasami - PAM vs AGNES

Rysunek 24 przedstawia wartości Adjusted Rand Index (ARI), który jest wskaźnikiem zewnętrznym, mierzącym zgodność między uzyskanym podziałem na klastry a rzeczywistą przynależnością obiektów do klas. Zarówno dla algorytmu PAM (linia niebieska), jak i AGNES (linia czerwona), wartości ARI są stosunkowo niskie. Dla PAM, wartości ARI wahają się od około 0.18 dla K=2, ze spadkiem do 0.12 dla K=4, a następnie delikatnym wzrostem do około 0.20 dla K=10 i K=11. Algorytm AGNES rozpoczyna z niemal zerowymi wartościami dla K=2, stopniowo rosnąc do około 0.28 dla K=10, K=11 i K=12. Niskie wartości ARI dla obu metod wskazują na ograniczoną zgodność z rzeczywistymi klasami w całym badanym zakresie K. Mimo że AGNES osiąga nieco wyższe wartości ARI dla większych K, różnice te nie są znaczące.

Podsumowując, wybór optymalnej liczby klastrów okazuje się niejednoznaczny. Indeks Silhouette sugeruje, że dla AGNES optymalna liczba skupień pod względem wewnętrznej spójności wynosi K=3 lub K=5 (bardziej K=5). Adjusted Rand Index natomiast nie dostarcza wyraźnego wskazania na optymalne K, a jego ogólnie niskie wartości świadczą o tym, że zarówno

PAM, jak i AGNES mają trudności w precyzyjnym odtworzeniu rzeczywistych klas zbioru danych Glass.

Żaden z algorytmów nie jest w stanie w zadowalającym stopniu odzwierciedlić rzeczywistej struktury danych. W porównaniu algorytmów, AGNES wykazuje nieznaczną przewagę nad PAM w aspekcie tworzenia wewnętrznie spójnych i dobrze oddzielonych skupień, co potwierdza indeks Silhouette. Jednakże, jeśli chodzi o zgodność z rzeczywistymi klasami, oba algorytmy wypadają podobnie i osiągają raczej niskie wyniki ARI. To ostatecznie sugeruje, że pomimo standaryzacji danych, rzeczywiste klasy w zbiorze Glass są ze sobą mocno przemieszane w przestrzeni cech, co utrudnia ich jednoznaczną separację za pomocą metod grupowania.

2.4 Interpretacja wyników grupowania - charakterystyki skupień

Jako optymalne K weźmy zatem 5, która wyróżnia się względnie wysokimi wartościami (jak przedstawiają Rysunki 24 oraz 23) na tle pozostałych wartości K.

Wyznaczono podziały na 5 skupień dla algorytmów PAM i AGNES (AGNES z metodą ‘complete’). Dane z przypisaniem do klastrów są dostępne w ramce danych ‘glass_clustered_K5’.

Z faktu, że wstawienie całej tabeli byłoby bardzo problematyczne, skonstruujmy nowy podział i wyświetlmmy tabele podsumowującą średnie wartości cech dla każdego klastra oddzielnie dla PAM i oddzielnie dla AGNES (complete) oraz w celu scharakteryzowania i rozróżnienia klastrów przeanalizujemy wykresy pudełkowe dla każdej zmiennej.

Tabela 15: Średnie wartości standaryzowanych cech dla każdego klastra (PAM, K=5)

Cluster_PAM	n_objects	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
1	24	1.32	0.63	0.60	-1.11	-1.32	-0.51	0.48	-0.15	-0.24
2	106	-0.37	-0.24	0.41	-0.01	0.14	0.29	-0.39	-0.29	-0.48
3	39	-0.14	-0.41	0.44	-0.24	0.15	0.05	-0.16	-0.30	1.69
4	18	1.78	-0.82	-1.63	-0.15	-0.40	-0.33	2.53	0.00	0.11
5	27	-0.69	1.52	-1.68	1.47	0.69	-0.56	-0.34	1.72	-0.44

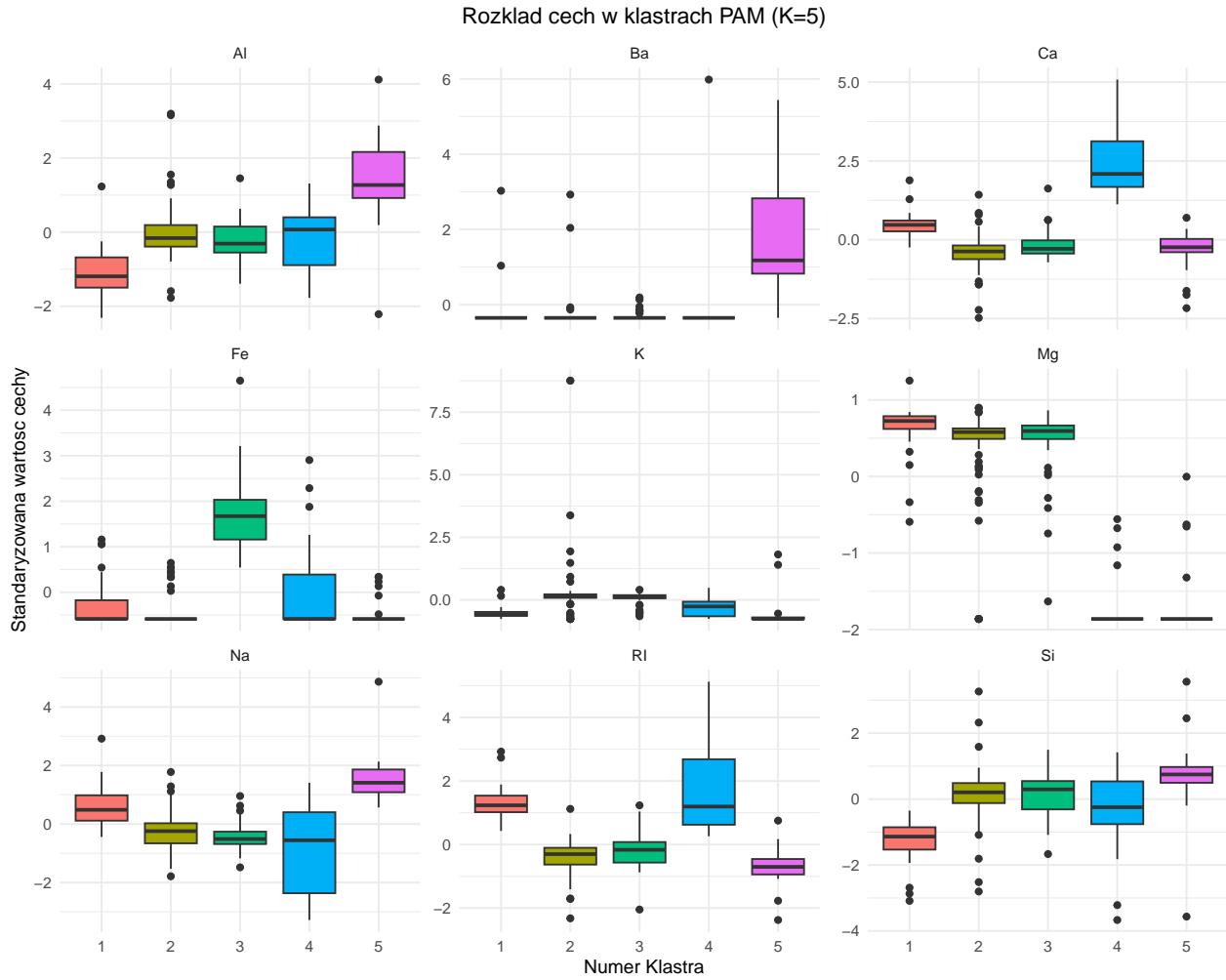
Medoidy (reprezentanci skupień) dla PAM (K=5) to obiekty o następujących indeksach: [1] “44” “43” “33” “171” “205”

Standaryzowane wartości cech dla medoidów (PAM, K=5):

Tabela 16: Standaryzowane wartości cech dla medoidów (PAM, K=5)

	Cluster	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
44	1	1.23	0.39	0.80	-1.45	-1.15	-0.50	0.55	-0.35	-0.59
43	2	-0.19	-0.24	0.49	-0.23	0.14	0.14	-0.26	-0.35	-0.59

33	3	-0.20	-0.68	0.55	-0.43	0.41	0.17	-0.28	-0.17	1.67
171	4	1.75	0.04	-1.86	0.27	-0.56	-0.27	2.31	-0.35	-0.59
205	5	-0.72	1.89	-1.86	1.65	0.84	-0.76	-0.17	1.00	-0.59



Rysunek 25: Wykresy pudełkowe dla cech w klastrach PAM

Klaster 1 - 24 obiekty: Wysoki indeks refrakcji (RI), podwyższone Na i Mg, bardzo niskie Al i Si. Medoid (obiekt 44) potwierdza tę charakterystykę ($RI = 1.23$, $Al = -1.45$, $Si = -1.15$). Może sugerować szkło specjalistyczne optyczne, dekoracyjne.

Klaster 2 - 106 obiektów: Największy klaster z cechami bliskimi średniej dla większości zmiennych. Medoid (obiekt 43) jest bardzo zbliżony do średnich wartości klastra ($RI = -0.19$, $Mg = 0.49$, $Al = -0.23$), wskazując na “typowe” szkło.

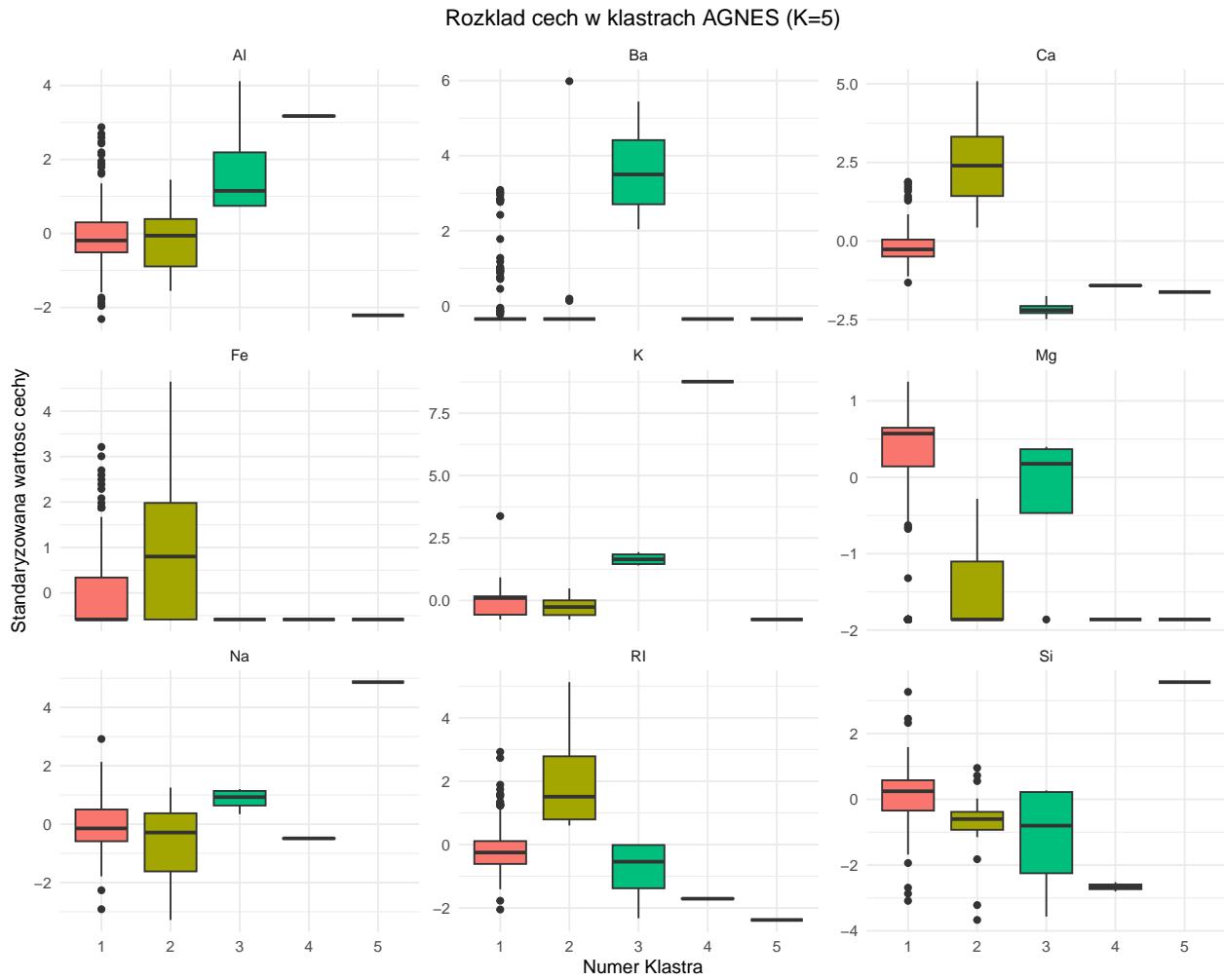
Klaster 3 - 39 obiektów: Wyróżnia się bardzo wysoką zawartością żelaza (Fe), podwyższonym Mg i Si, oraz niskim Na. Medoid (obiekt 33) potwierdza wysoką zawartość żelaza (Fe = 1.67) i magnezu (Mg = 0.55), z niskim Na (-0.68), sugerując, że może to odpowiadać szkłu zabarwionemu (żelazo to częsty barwnik do szkła).

Klaster 4 - 8 obiektów: Bardzo wysoki indeks refrakcji (RI) i wapń (Ca), bardzo niskie Mg. Medoid (obiekt 171) potwierdza wysokie RI (1.75) i Ca (2.31), z bardzo niskim Mg (-1.86). Wskazuje na szkło optyczne (okulary).

Klaster 5 - 27 obiektów: Wysoka zawartość baru (Ba), sodu (Na), glinu (Al) i krzemu (Si), przy bardzo niskim magnezje (Mg). Medoid (obiekt 205) potwierdza te cechy (Ba = 1.00, Na = 1.89, Al = 1.65, Mg = -1.86).

Tabela 17: Średnie wartości standaryzowanych cech dla każdego klastra (AGNES, K=5)

Cluster_AGNES	n_objects	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
1	191	-0.12	0.02	0.16	-0.05	0.10	-0.10	-0.13	-0.08	-0.06
2	16	1.97	-0.71	-1.49	-0.10	-0.80	-0.27	2.43	0.11	0.92
3	4	-0.85	0.85	-0.28	1.79	-1.22	1.66	-2.16	3.62	-0.59
4	2	-1.71	-0.49	-1.86	3.17	-2.66	8.76	-1.41	-0.35	-0.59
5	1	-2.38	4.86	-1.86	-2.21	3.56	-0.76	-1.62	-0.35	-0.59



Rysunek 26: Wykresy pułapkowe dla cech w klastrach AGNES

Klaster 1 - 191 obiektów: to największy klaster, obejmujący większość próbek. Cechy są bliskie średniej ($RI = -0.12$, $Na = 0.02$, $Mg = 0.16$), co sprawia, że jest bardzo heterogeniczny i pozabawiony wyraźnych wyróżników. Możemy zasugerować, że reprezentuje on pospolite szkło (np. okienne), a jego dominacja sugeruje trudności AGNES w dokładnej separacji danych.

Klaster 2 - 16 obiektów: charakteryzuje się bardzo wysokim indeksem refrakcji (RI) i wapniem (Ca), podwyższonym żelazem (Fe) oraz niskim magnezem (Mg) i krzemem (Si). To jednorodna grupa, która prawdopodobnie odpowiada szkłu optycznemu lub zabarwionemu.

Klaster 3 - 4 obiekty: ma bardzo wysoką zawartość baru (Ba), glinu (Al) i potasu (K), przy ekstremalnie niskim wapniu (Ca) i krzemie (Si).

Klaster 4 - 2 obiekty: wyróżnia się ekstremalnie wysokim potasem (K) i glinem (Al), a bardzo niskim krzemem (Si), magnezem (Mg) i indeksem refrakcji (RI).

Klaster 5 - To pojedynczy obiekt-outlier o ekstremalnie wysokim sodzie (Na) i krzemie (Si), ale bardzo niskim indeksem refrakcji (RI) i glinie (Al).

Szklka wchodzące w skład Klastra 3,4 i 5 pod względem ilości obiektów nazwalibyśmy szkłem bardzo rzadkim oraz specjalistycznym (outliery), któremu ciężko jest przypisać jakiekolwiek zastosowanie, czy użyteczność.

Podsumowując, PAM jest bardziej skuteczny w identyfikacji zróżnicowanych typów szkła, tworząc zrównoważone klastry z wyraźnymi cechami chemicznymi, co czyni go lepszym wyborem do analizy różnorodności w zbiorze „Glass”. AGNES z metodą „complete” lepiej radzi sobie z wychwytywaniem nietypowych obiektów, takich jak rzadkie typy szkła czy outliery, ale jego tendencja do grupowania większości danych w jeden ogólny klaster zmniejsza zdolność do separacji danych. Oba algorytmy potwierdzają, że zbiór „Glass” jest trudny do klastrowania ze względu na słabą separowalność klas, co wynika z niskich wartości ARI i przemieszania cech w przestrzeni danych, utrudniając precyzyjne odtworzenie rzeczywistej struktury klas.