Raport 4

Eksploracja danych

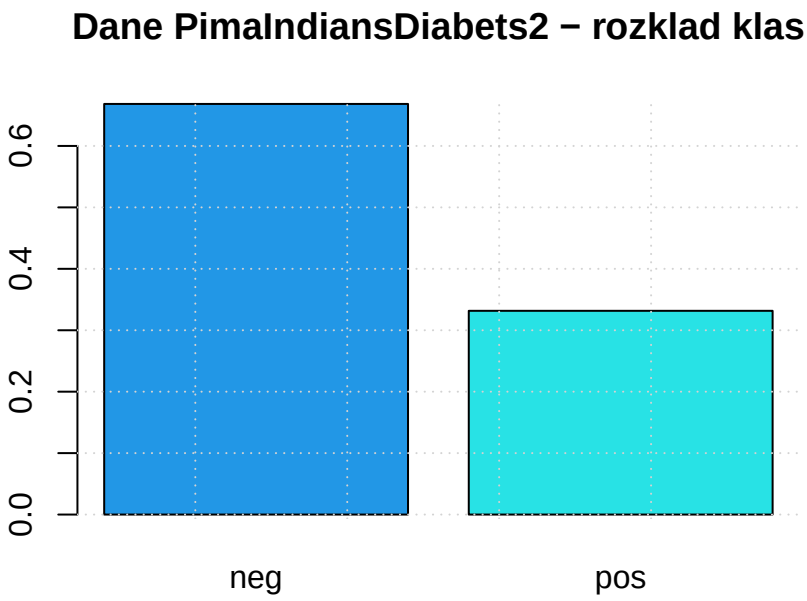
Olga Foriasz 277529, Szymon Smoła 282252

2025-06-18

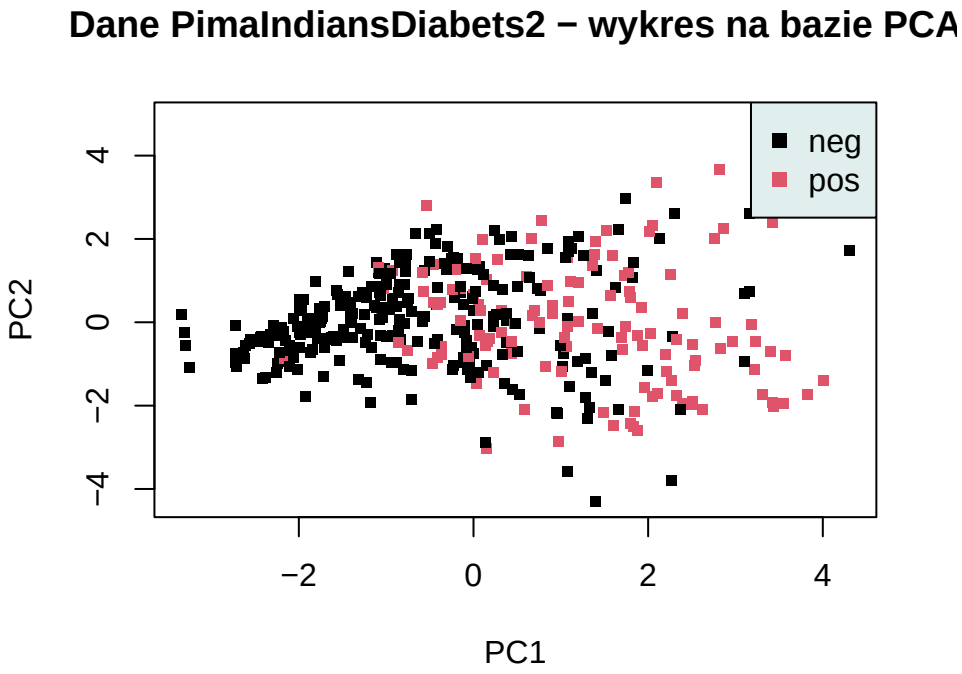
Spis treści

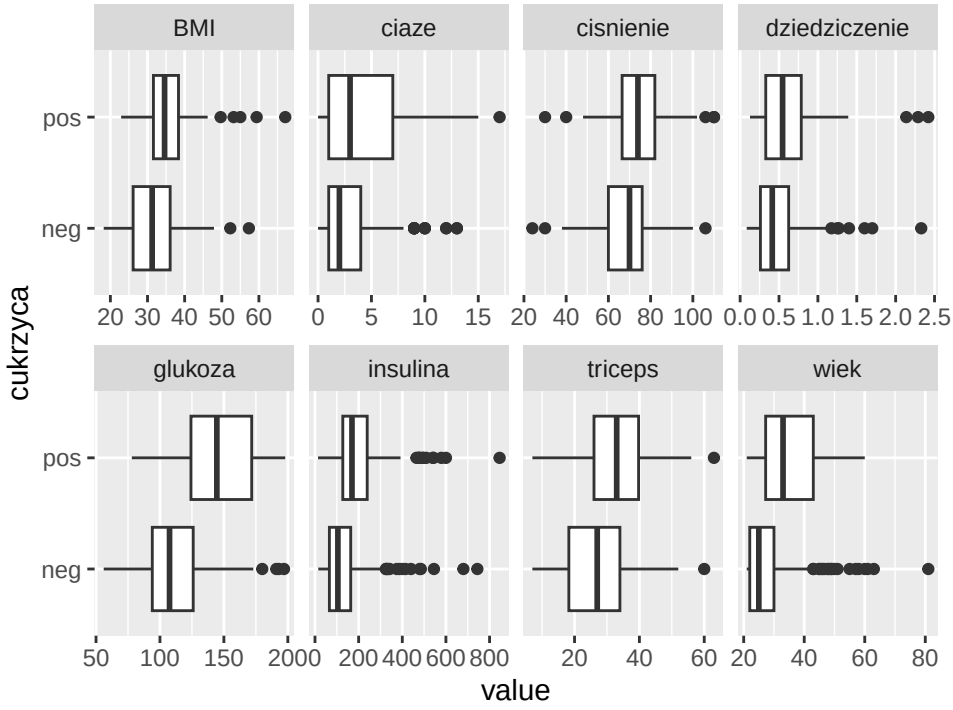
1 Zaawansowane metody klasyfikacji

1.1 Rodziny klasyfikatorów/uczenie zespołowe

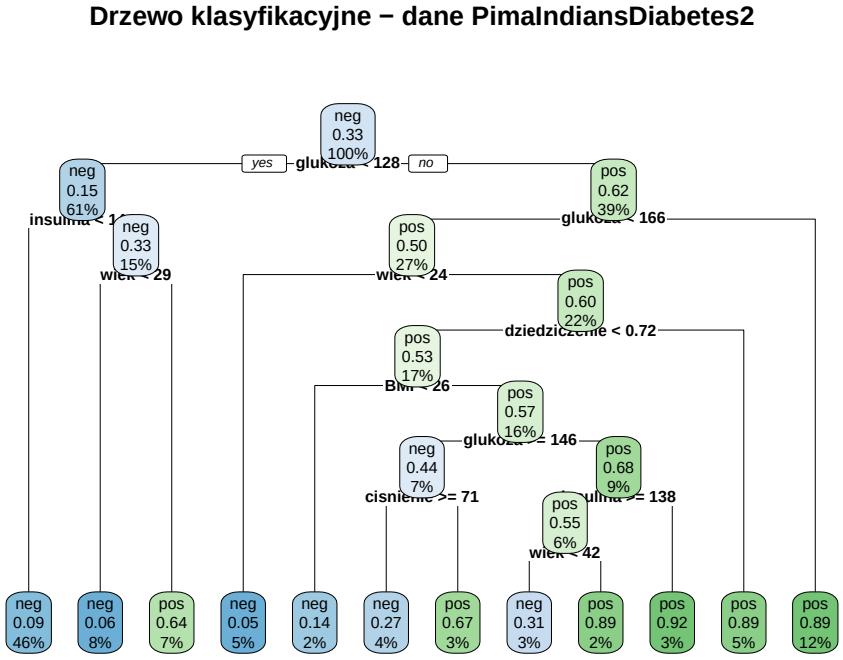



```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.5999 1.2477 1.0949 0.9776 0.84863 0.63358 0.55774
## Proportion of Variance 0.3199 0.1946 0.1499 0.1195 0.09002 0.05018 0.03888
## Cumulative Proportion  0.3199 0.5145 0.6644 0.7839 0.87387 0.92404 0.96293
##                           PC8
## Standard deviation     0.54458
## Proportion of Variance 0.03707
## Cumulative Proportion  1.00000
```


Dane PimaIndiansDiabets2 − wykres na bazie PCA
neg
pos
PC2
PC1
4
2
0
−2
−4
−2
0
2
4


BMI
ciaze
cisnienie
dziedziczenie
glukoza
insulina
triceps
wiek
pos
neg
cukrzyca
value
20 30 40 50 60
0 5 10 15
20 40 60 80 100
0.0 0.5 1.0 1.5 2.0 2.5
50 100 150 200
0 200 400 600 800
20 40 60
20 40 60 80

1.1.1 Pojedyńcze drzewo klasyfikacyjne



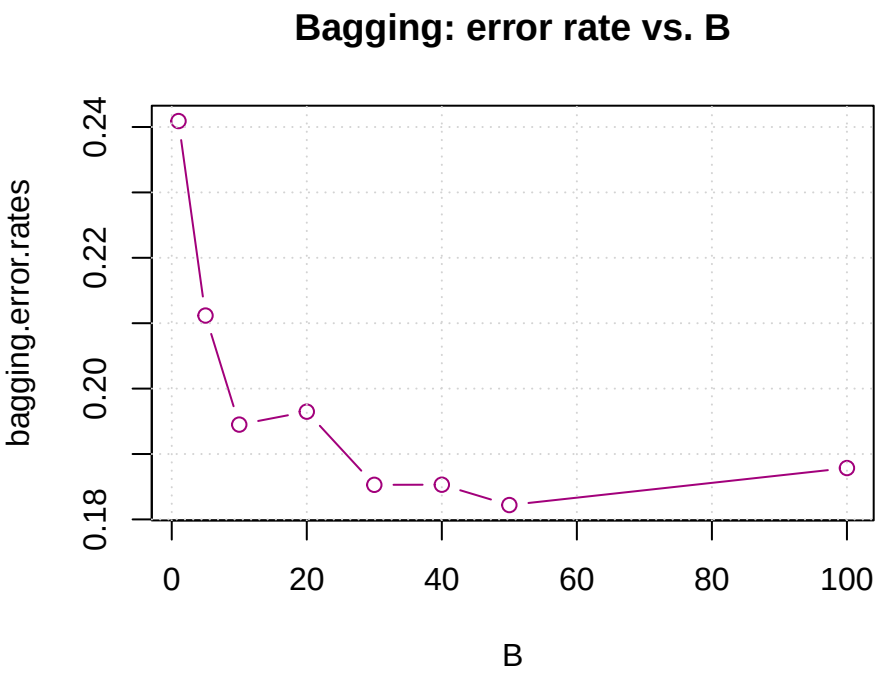
Rysunek 1: Drzewo klasyfikacyjne

Na powyższym Rysunku 1 przedstawione zostało drzewo klasyfikacyjne dla całego zbioru danych. Poniżej znajdują się dane ile wynosiły błędy klasyfikacyjne dla zbioru testowego:

```
## Błąd klasyfikacji - zbiór uczący: 0.111
```

```
## Błąd klasyfikacji - zbiór testowy: 0.244
```

1.1.2 Metoda bagging



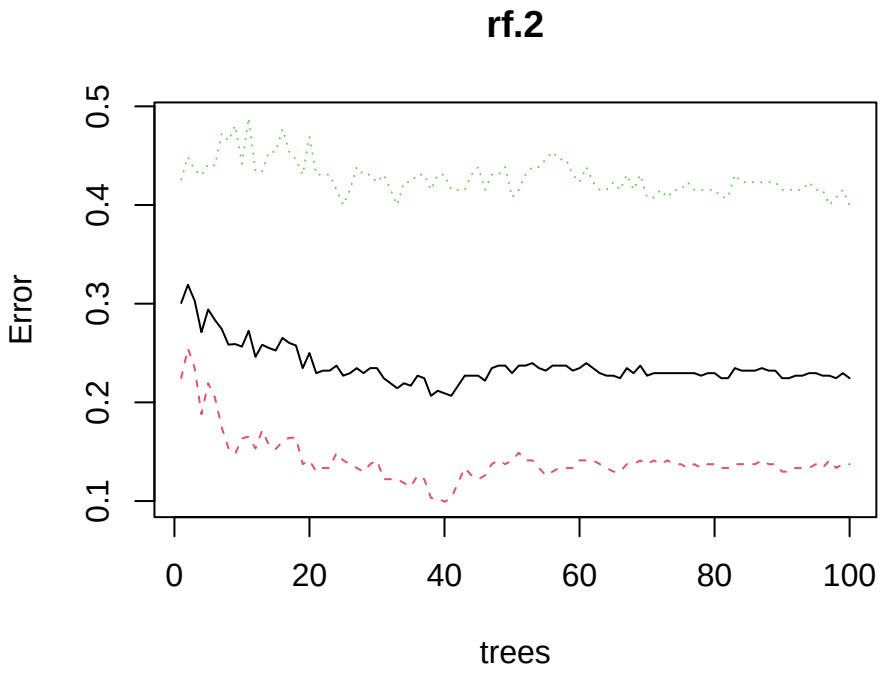
Rysunek 2: Wykresy dla metody bagging

Można zauważyć na Rysunku 2, że wraz ze wzrostem liczby drzew błąd wyraźnie maleje do około 40 drzew, po czym stabilizuje się i przestaje znacząco spadać. Dalsze zwiększanie liczby drzew (np. do 100) nie poprawia wyników, a wręcz może nieznacznie je pogorszyć. Oznacza to, że optymalna liczba drzew mieści się w przedziale 40–60.

1.1.3 Metoda Random Forest

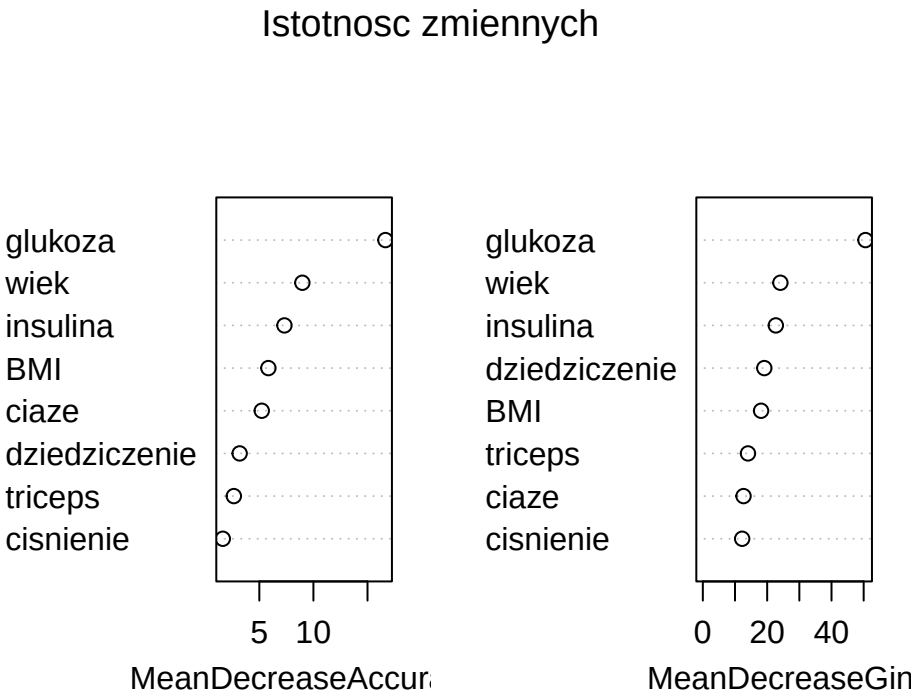
```
##             real.labels
## pred.labels neg pos
##         neg 262   1
##         pos   0 129
```

```
##     neg pos class.error
## neg 226  36   0.1374046
## pos  52  78   0.4000000
```



Rysunek 3: Wykres dla metody Random Forest

Wykres błędu dla metody Random Forest wskazuje, że model osiąga stabilność po około 40 drzewach — dalsze zwiększanie ich liczby nie przynosi istotnej poprawy wyników. Model lepiej klasyfikuje osoby bez cukrzycy, natomiast gorzej radzi sobie z wykrywaniem przypadków choroby. Taka asymetria może wynikać z nierównomiernego rozkładu klas w danych treningowych. Warto rozważyć zastosowanie technik wyrównania liczebności klas (np. oversampling, undersampling) lub innych miar oceny skuteczności modelu, które lepiej uwzględniają nierównowagę klas.



Rysunek 4: Istotność zmiennych

Na wykresie nr 4 możemy zauważyć jakie zmienne odgrywają największą rolę w przyporządkowywaniu próby do zbioru. W obu przypadkach najbardziej znacząca jest zmienna glukoza. Widać także, że zdecydowanie dominuje w porównaniu do innych zmiennych.

1.1.4 Wnioski

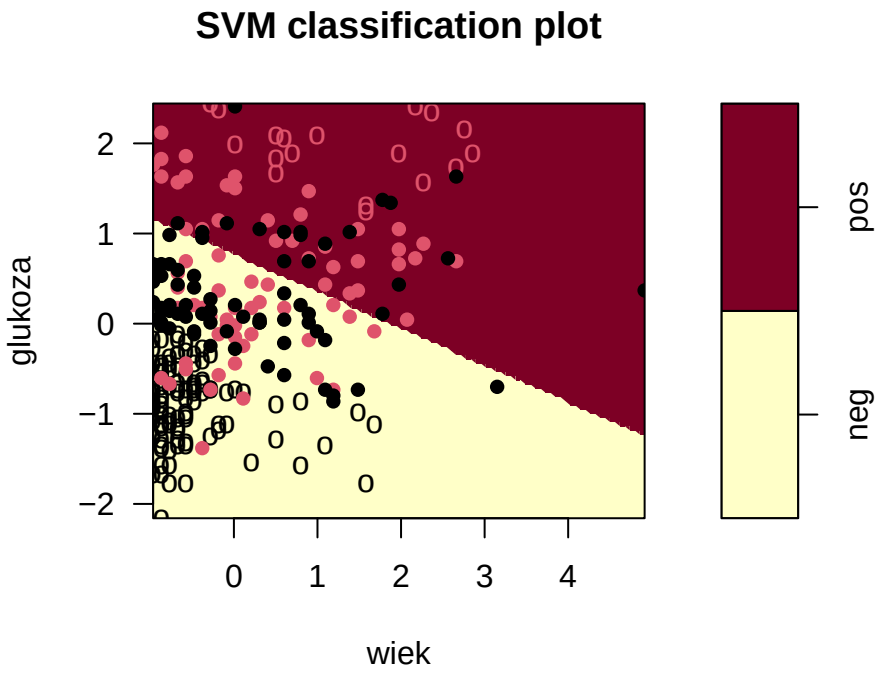
Przy zastosowaniu pojedynczego drzewa klasyfikacyjnego, błąd klasyfikacji dla zbioru testowego wynosił 24,4%. Zastosowanie metody bagging przyniosło istotną poprawę — przy odpowiednim doborze liczby replikacji B = 40, błąd klasyfikacyjny spadł do około 18%. Oznacza to wyraźną redukcję błędu w porównaniu do pojedynczego drzewa.

W przypadku metody Random Forest zaobserwowano, że błąd przypisania do jednej z klas był znacznie niższy niż do drugiej — różnica wynosiła około 30 punktów procentowych. Całkowity błąd klasyfikacji od momentu ustalenia liczby drzew na poziomie 20 stabilizował się i utrzymywał na poziomie nieco ponad 20%.

Podsumowując, najlepsze wyniki uzyskano dzięki metodzie bagging, która pozwoliła na największą redukcję błędu względem pojedynczego drzewa decyzyjnego. Choć metoda Random Forest również poprawiła jakość klasyfikacji, to jej skuteczność była mniejsza niż w przypadku baggingu — różnica w błędzie klasyfikacji między tymi dwiema metodami wynosiła około 10 punktów procentowych na korzyść baggingu.

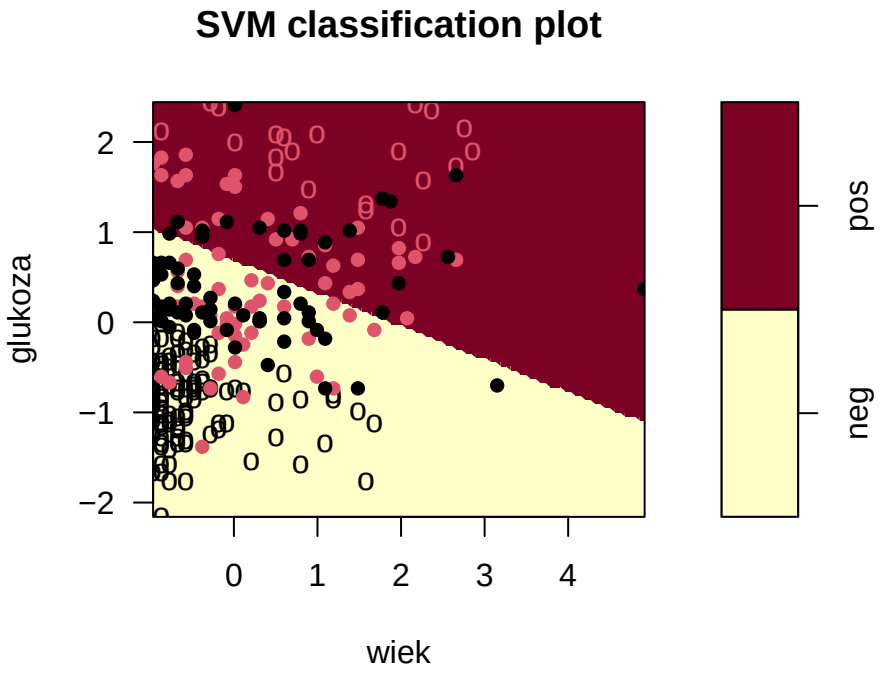
1.2 Metoda wektorów nośnych (SVM)

1.2.1 Porównanie skuteczności funkcji jądrowych i parametru kosztów



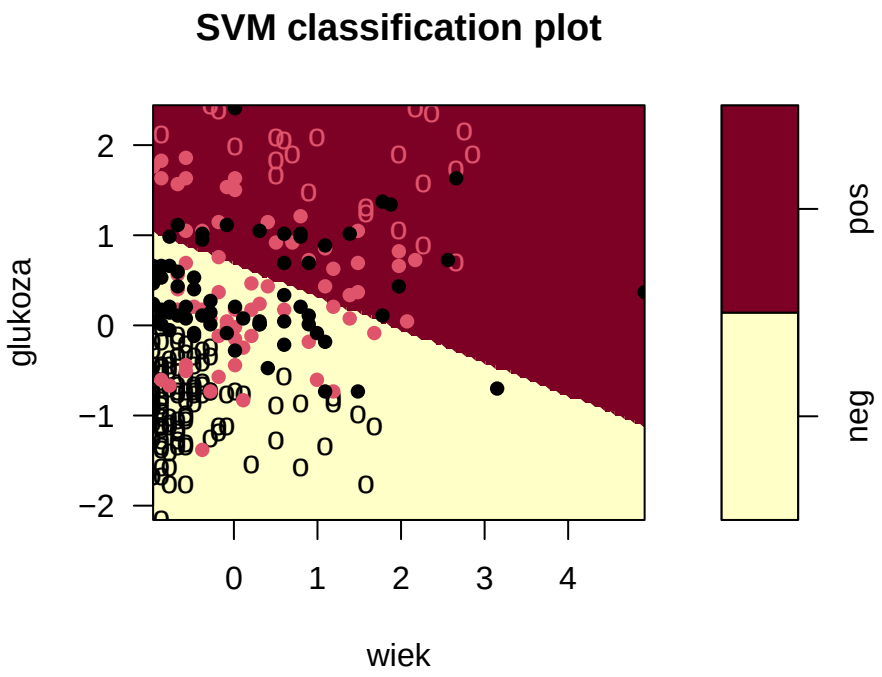
Rysunek 5: Wykresy funkcji jądrowych - jądro liniowe, C=0.1

```
## Skuteczność klasyfikacji: 0.802
```



Rysunek 6: Wykresy funkcji jądrowych - jądro liniowe, C=1

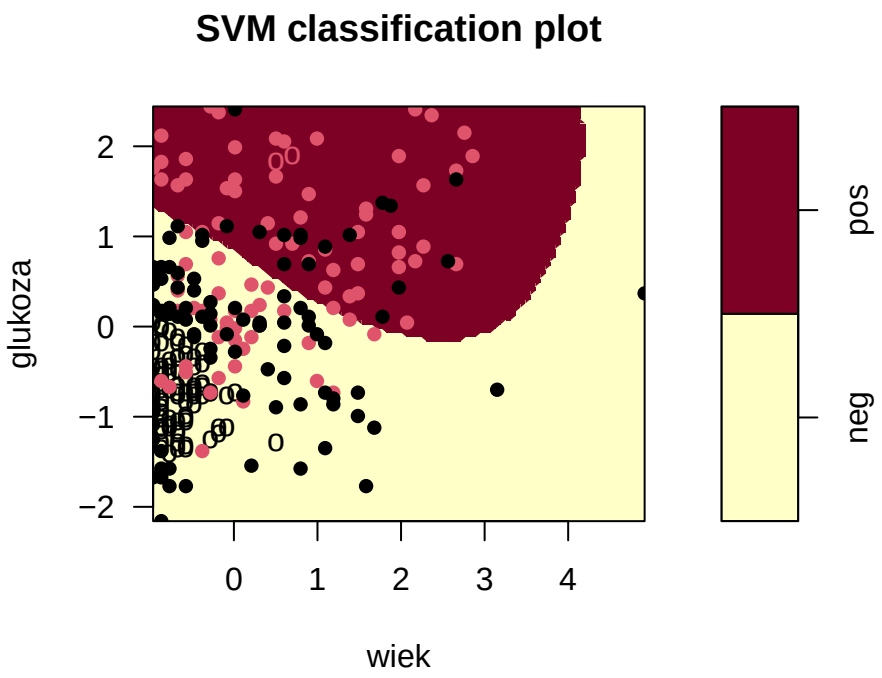
```
## Skuteczność klasyfikacji: 0.786
```



Rysunek 7: Wykresy funkcji jądrowych - jądro liniowe, C=10

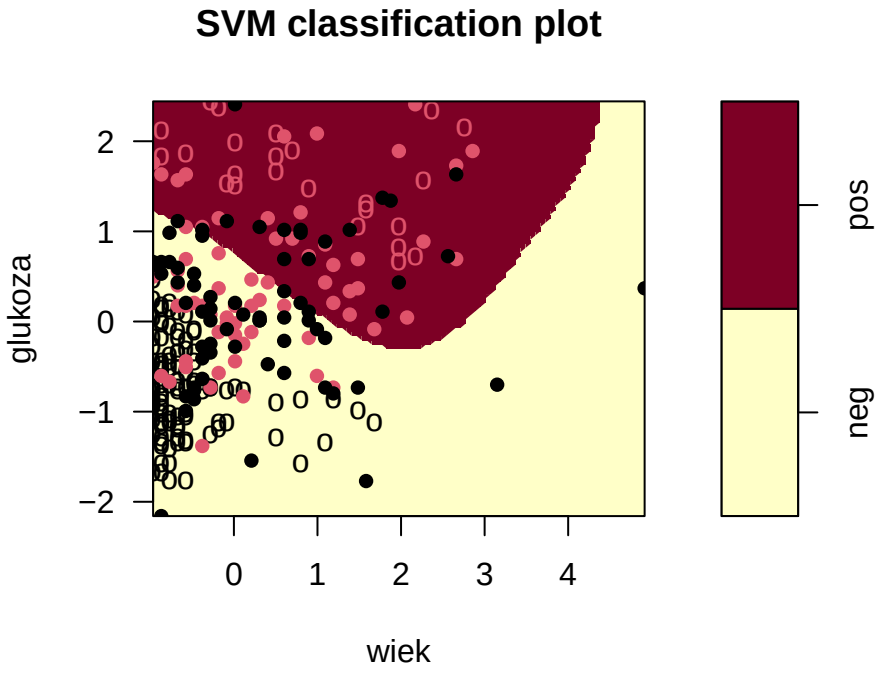
```
## Skuteczność klasyfikacji: 0.786
```

Na wykresach 5–7 przedstawiono funkcje decyzyjne SVM dla zmiennych glukoza oraz wiek, z wykorzystaniem jądra liniowego i różnych wartości parametru kosztu. Najwyższą skuteczność klasyfikacji uzyskano przy najmniejszej wartości parametru C=0.1 Różnice w skuteczności dla wyższych wartości parametru są jednak niewielkie i wynoszą około 2%.



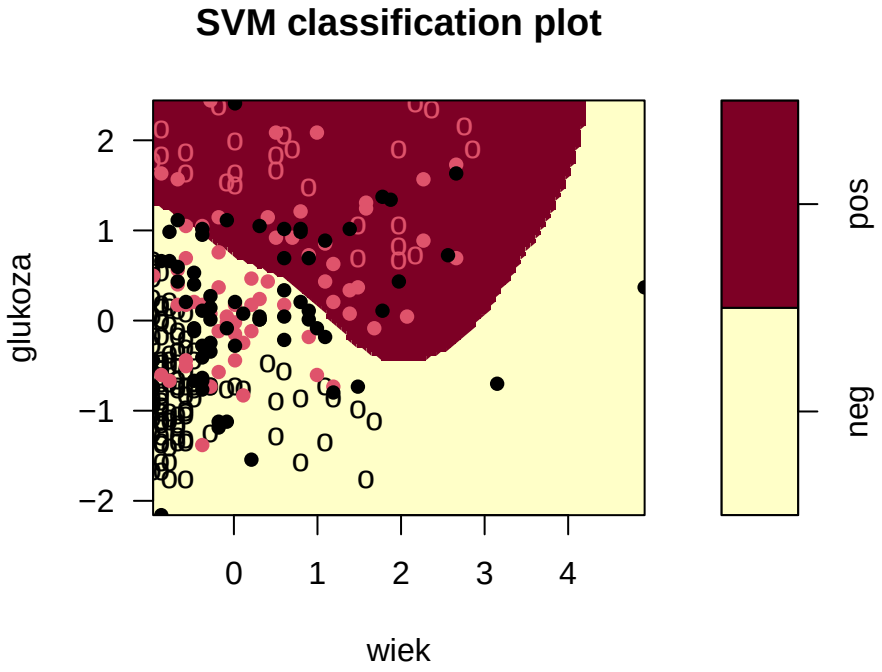
Rysunek 8: Wykresy funkcji jądrowych - jądro radialne, C=0.1

```
## Skuteczność klasyfikacji: 0.817
```



Rysunek 9: Wykresy funkcji jądrowych - jądro radialne, C=1

```
## Skuteczność klasyfikacji: 0.794
```



Rysunek 10: Wykresy funkcji jądrowych - jądro radialne, C=10

```
## Skuteczność klasyfikacji: 0.802
```

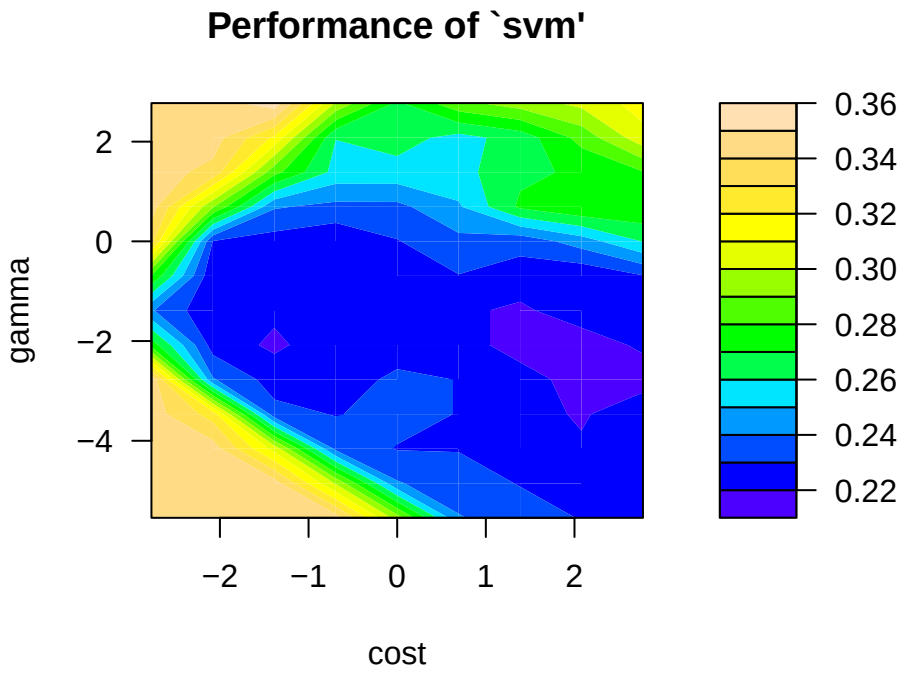
Na wykresach 8–10 przedstawiono funkcje decyzyjne SVM dla zmiennych glukoza oraz wiek, z wykorzystaniem jądra radialnego i różnych wartości parametru kosztu C. Najlepszą skuteczność klasyfikacji uzyskano dla C = 0.1 – wyniosła 81,7%. Niższą skuteczność (80,2%) osiągnięto przy C = 10. Natomiast jeszcze gorszy wynik (79,4%) uzyskano przy C = 1.

Wybór funkcji jądrowej oraz parametru kosztu C wpływa na skuteczność klasyfikacji. Różnice są jednak niewielkie, a przy odpowiednio dobranych wartościach parametru C wyniki mogą być porównywalne. Niemniej jednak, dla tych samych parametrów funkcje jądrowe radialne osiągnęły lepsze rezultaty.

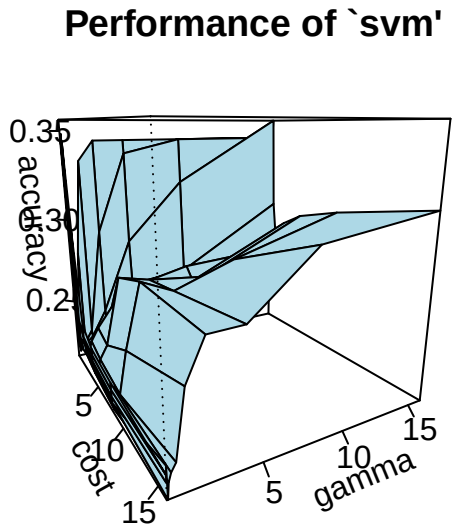
1.2.2 Dobranie najlepszych parametrów dla jądra radialnego

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
##    16 0.0625
##
## - best performance: 0.214245
```



Rysunek 11: Wykresy funkcji jądrowej (jądro radialne)



Rysunek 12: Wykresy funkcji jądrowej (jądro radialne)

```
## Skuteczność klasyfikacji: 0.802
```

Porównując dokładność klasyfikacji dla modelu z optymalnie dobranymi parametrami (gamma = 0.625 oraz C = 16) z modelem skonstruowanym dla domyślnych parametrów (C = 1), otrzymano taką samą skuteczność – 80,2%. Co ciekawe, jeszcze lepszy wynik – 80,9% – uzyskano dla modelu z niższą wartością parametru C = 0.1, bez przeprowadzania pełnej optymalizacji. Oznacza to, że w tym przypadku optymalizacja parametrów nie przyniosła poprawy skuteczności klasyfikatora, a prostszy model poradził sobie najlepiej.

1.3 Porównanie skuteczności metod

Porównując metody klasyfikatorów z rodziny drzew decyzyjnych oraz uczenia zespołowego (punkt a) z metodą wektorów nośnych SVM (punkt b), można zauważyć, że wyższą skuteczność w większości przypadków uzyskano przy wykorzystaniu metody SVM, mimo że w analizie tej ograniczono się jedynie do dwóch zmiennych: glukozy i wieku.

Najniższą skuteczność w ramach SVM odnotowano przy zastosowaniu jądra liniowego – 74,8%, co i tak jest porównywalne lub lepsze niż wyniki uzyskane w niektórych wariantach metod klasyfikatorów zespołowych. Najlepszy wynik osiągnięto dla jądra radialnego z parametrem C=0.1, gdzie skuteczność klasyfikacji wyniosła 81,7%.

W przypadku uczenia zespołowego, najlepsze rezultaty uzyskano stosując metodę bagging przy liczbie replikacji równej 40 – błąd klasyfikacji wyniósł wtedy 18%, co odpowiada skuteczności 82%. Dla pozostałych wariantów błąd klasyfikacji był wyraźnie wyższy, wynosząc ok. 25%.

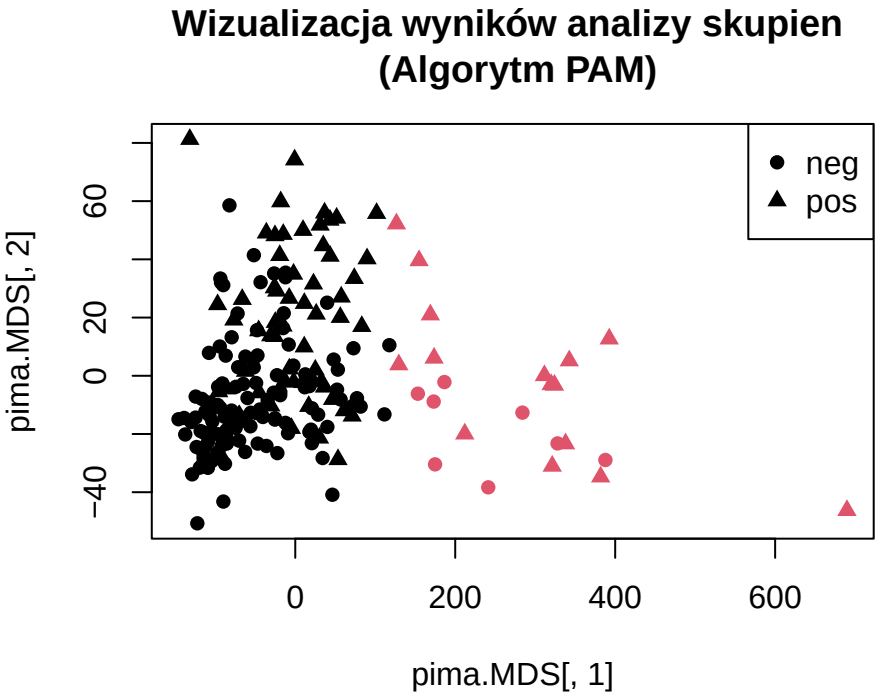
Podsumowując, najlepszą skuteczność uzyskano przy metodzie bagging (82%), jednak metoda wektorów nośnych również wykazała bardzo wysoką, zbliżoną skuteczność, mimo uproszczonego podejścia. SVM okazała się stabilną i skuteczną metodą klasyfikacyjną, szczególnie biorąc pod uwagę, że nie wykorzystano wszystkich dostępnych zmiennych predykcyjnych.

2 Analiza skupień – algorytmy grupujące i hierarchiczne

2.1 Wybór i przygotowanie danych

Wylosowano 200 rekordów z danych PimaIndiansDiabetes2 z pakietu mlbench. Dane te mają jedną zmienną grupującą nasze obserwacje na dwie klasy (cukrzyków i osoby bez cukrzycy). Standaryzacja była konieczna ponieważ cechy ilościowe mają różne jednostki i zakresy wartości.

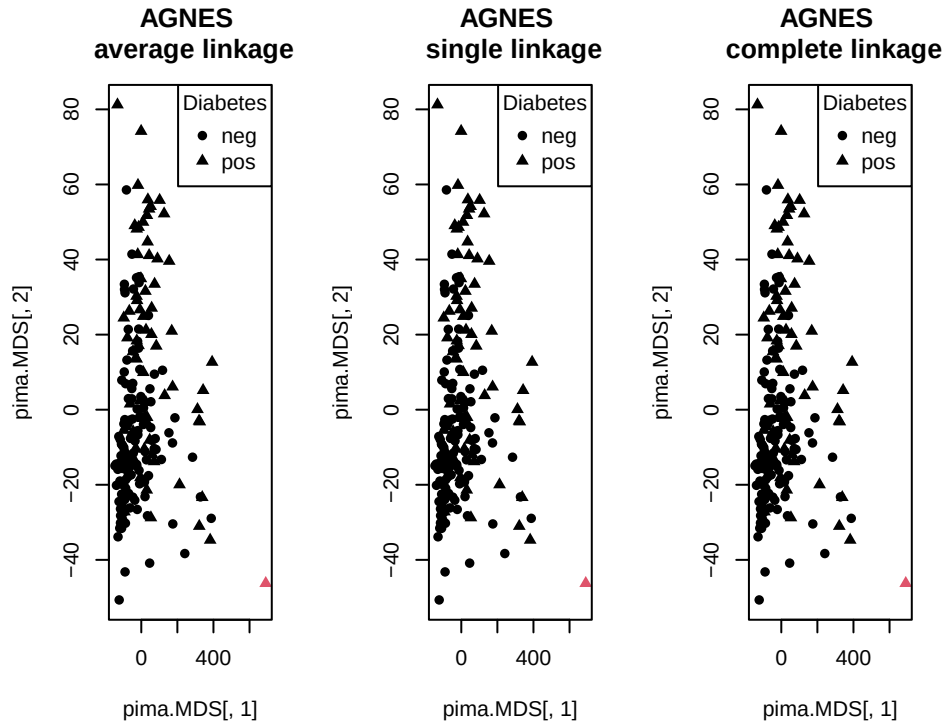
2.2 Wizualizacja wyników grupowania



Rysunek 13: Analiza skupień - PAM

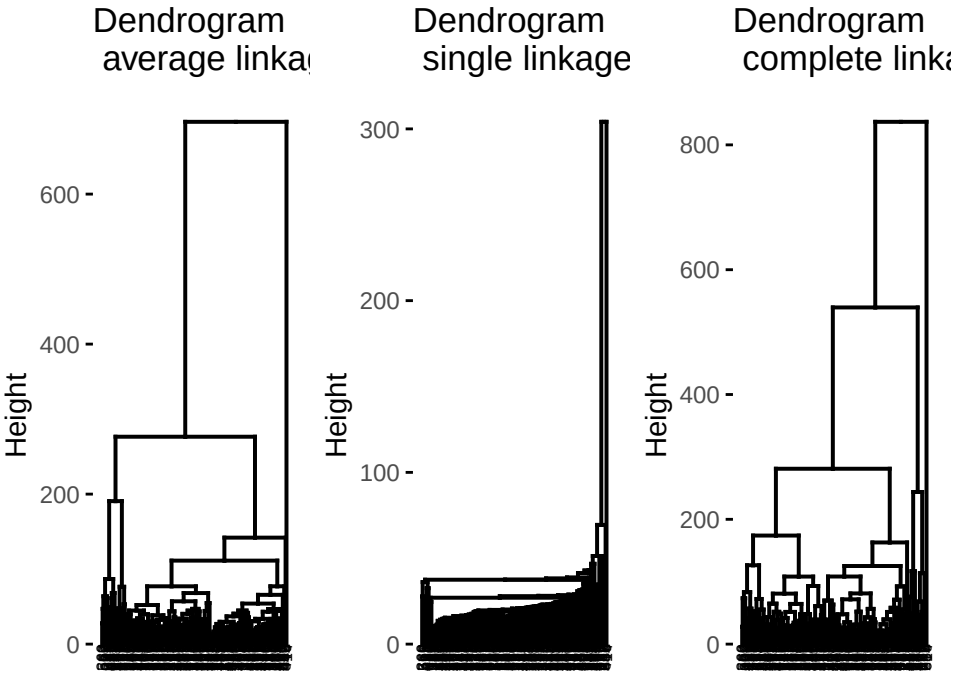
Wyniki grupowania algorytmem **PAM** dla **K=2** z rysunku 13 słabo pokazują zgodność z rzeczywistymi klasami. Wykres MDS pozwala zauważyć, że:

- Grupy są względnie zwarte, ale nie idealnie separowalne
- Istnieje niewielka korelacja między przynależnością do skupień a rzeczywistymi klasami



Rysunek 14: Analiza skupień - AGNES

Wykresy przedstawione na rysunku 14 dla ustalonego **K=2** są niejako zwarte i spójne, ale widać na nich sporą separację przestrzenną. Podział na klastry metodami AGNES ma w naszym przypadku małą zgodnośc z rzeczywistą przynależnością obiektów do klas.



Rysunek 15: Dendogramy - algorytm AGNES

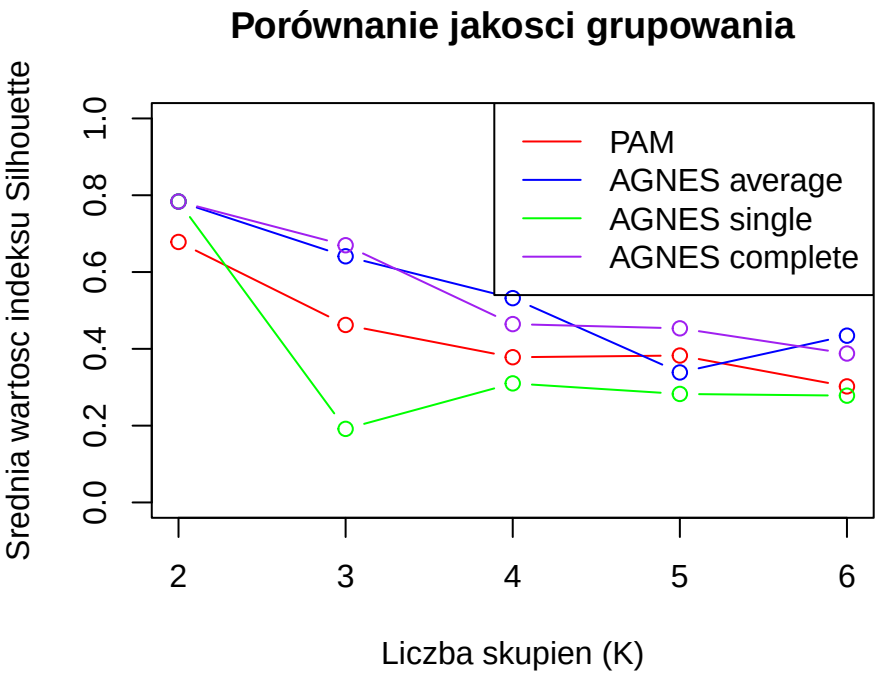
Na rysunku 15 porównano trzy metody łączenia skupień:

- Average linkage: tworzy całkiem zrównoważone grupy
- Single linkage: wykazuje tendencję do tworzenia długich, łańcuchowych skupień
- Complete linkage: tworzy bardziej zwarte skupienia

Dendrogramy wyraźnie pokazują różnice w strukturze hierarchicznej w zależności od metody łączenia. Metody average i complete wydają się być najlepiej zbilansowane, a metoda single utworzyła wąskie i wydłużone skupiska.

2.3 Ocena jakości grupowania

2.3.1 Wskaźniki wewnętrzne



Rysunek 16: Porównanie jakości grupowania - Silhouette

Analiza średnich wartości indeksu Silhouette dla różnych metod i liczby skupień K=2-6 na rysunku 16 pokazuje, że:

- Najwyższe wartości osiągają algorytmy AGNES dla K=2 (około 0.8)
- Metody hierarchiczne osiągają podobne wyniki, z niewielką przewagą metody average
- Optymalna liczba skupień to K=2 dla wszystkich metod

2.3.2 Wskaźniki zewnętrzne

Porównanie z rzeczywistymi etykietami klas (diabetes):

PAM: 69% zgodności

AGNES complete: 63% zgodności

AGNES single: 63% zgodności

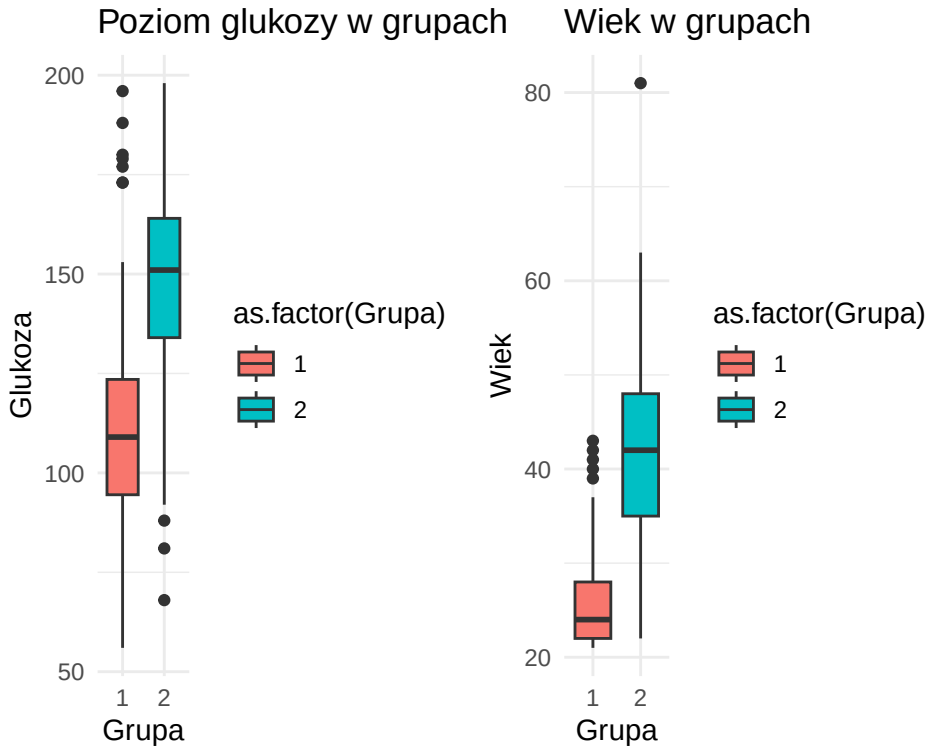
AGNES average: 63% zgodności

Algorytm PAM osiągnął najlepszą zgodność z rzeczywistym podziałem na klasy.

2.4 Interpretacja wyników grupowania

Tabela 1: Średnie wartości cech w skupieniach

Grupa	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age
1	1.91	111.11	68.75	27.41	122.29	32.23	0.58	25.88
2	6.16	147.65	75.57	32.78	224.91	34.93	0.43	41.77



Rysunek 17: Porównanie wykresów pudełkowych

Analiza charakterystyk skupień dla **K=2** przedstawionych na rysunku 17 pokazuje znaczące różnice między grupami:

Grupa 1: niższe średnie wartości glukozy, niższy wiek, mniejsze BMI

Grupa 2: wyższe wartości glukozy, starsi pacjenci, wyższe BMI

Tabela 2: Wartości cech - medoidy

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
86	2	110	74	29	125	32.4	0.70	27	1
646	2	157	74	35	440	39.4	0.13	30	1

Medoidy (reprezentanci skupień) dla algorytmu PAM:

Grupa 1: pacjentka z umiarkowanymi wartościami glukozy (110), BMI (32.4), młodsza (27 lat)

Grupa 2: pacjentka z podwyższoną glukozą (157), wyższym BMI (39.4), starsza (30 lat)