Raport 4

Eksploracja danych

Olaf Masłowski, album 277543

2025-07-09

Spis treści

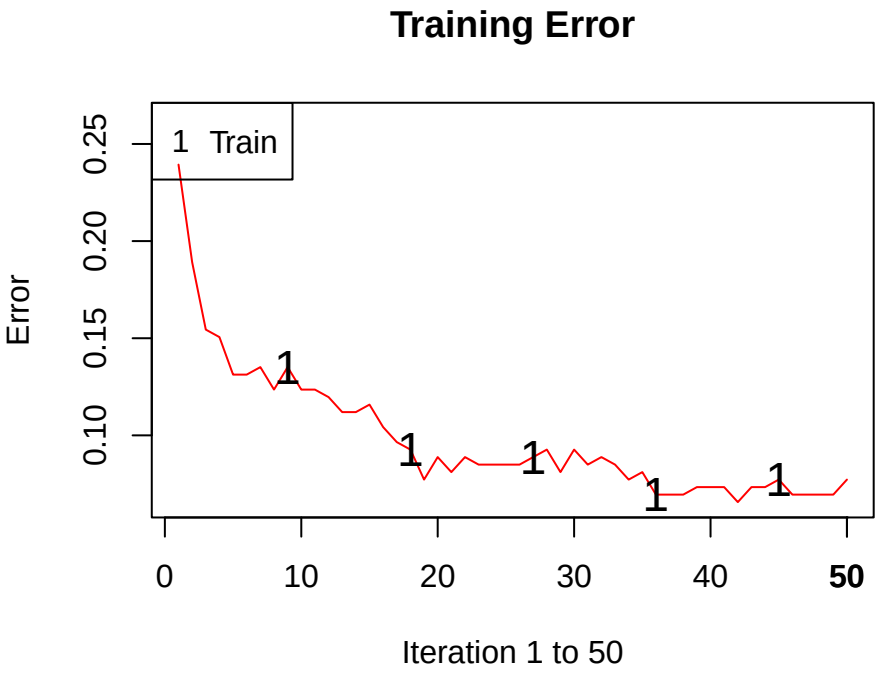
1 Wstęp

W pierwszej części będę kontynuować analizę danych PimaIndiansDiabetes wykorzystując rodziny klasyfikatorów.

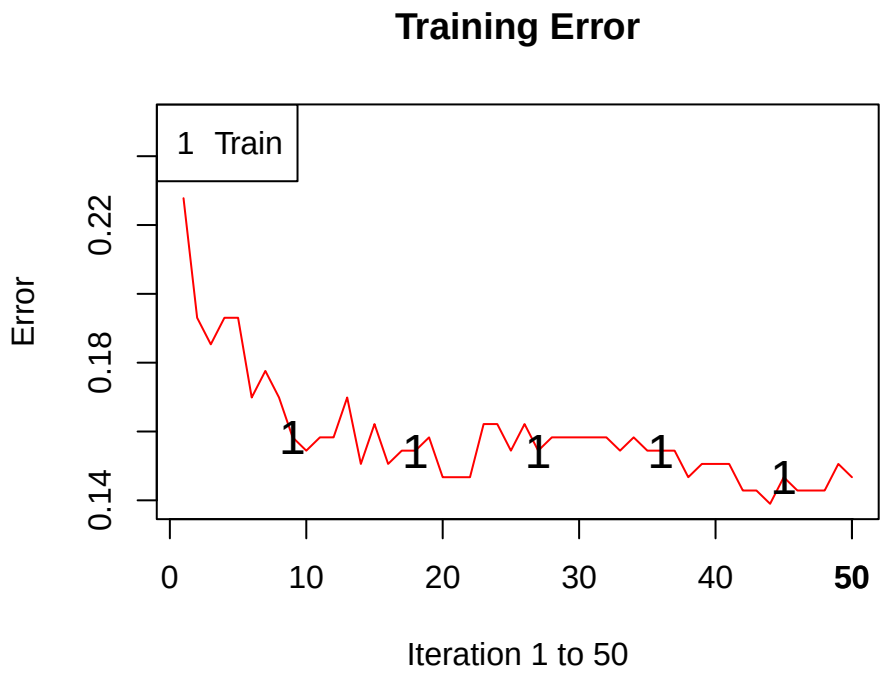
2 Zadanie 1 - Zaawansowane metody klasyfikacji

2.1 Boosting

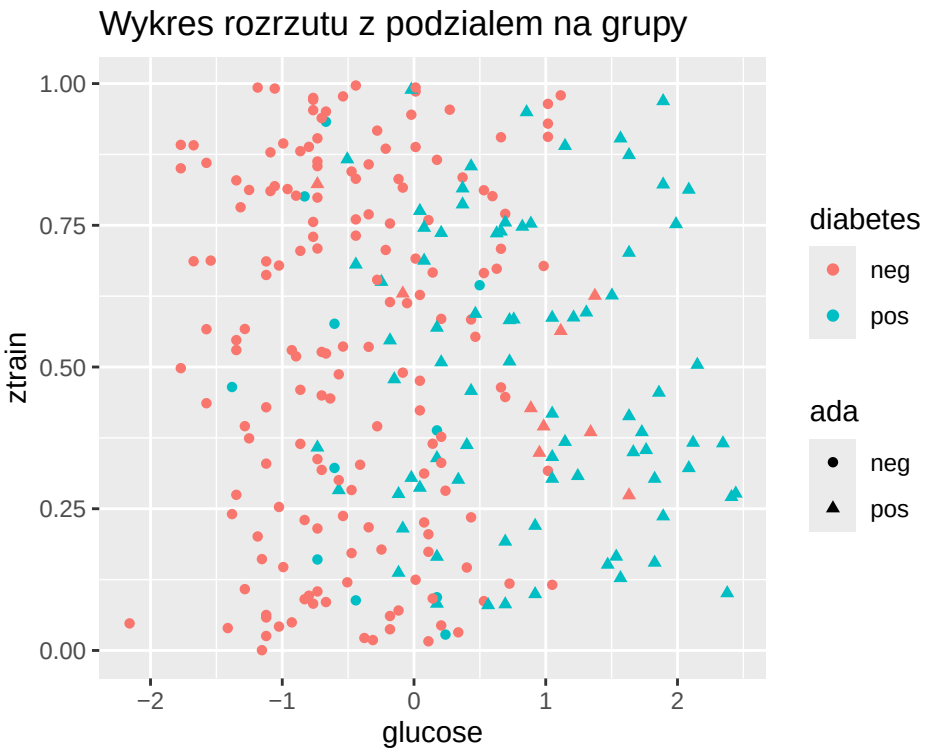
Wykorzystam funkcję ada(){ada} do implementacji algorytmu metody boosting - Adaboost.



Rysunek 1: Błąd klasyfikacji na zbiorze treningowym w zależności od liczby iteracji. Wszystkie zmienne



Rysunek 2: Błąd klasyfikacji na zbiorze treningowym w zależności od liczby iteracji. Zmienne glucose i age



Rysunek 3: Wykres rozrzutu z podziałem na wynik klasyfikacji i rzeczywistą etykietkę dla danych uczących z wykorzystaniem wszystkich zmiennych. Liczba iteracji - 50.

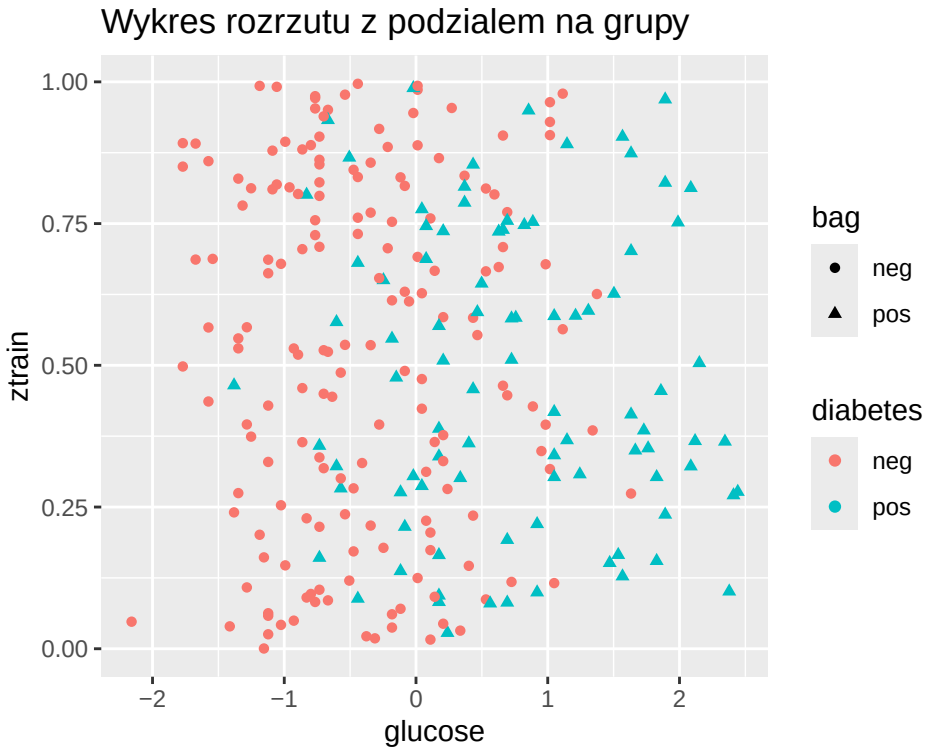
	parametry	error
1	train, iter = 20	0.104
2	train, iter = 50	0.077
3	train, iter = 20, g + a	0.154
4	train, iter = 50, g + a	0.147
5	test, iter = 20	0.203
6	test, iter = 50	0.226
7	test, iter = 20, g + a	0.241
8	test, iter = 50, g + a	0.271
9	crossvalidation, iter=50, g + a	0.232
10	crossvalidation, iter = 50	0.212

Tabela 1: Zestawienie błądów klasyfikacji

2.2 Bagging

Bagging classification trees with 25 bootstrap replications

Call: bagging.data.frame(formula = diabetes ~ ., data = trainset, nbag = 25, minsplit = 0, cp = 0)

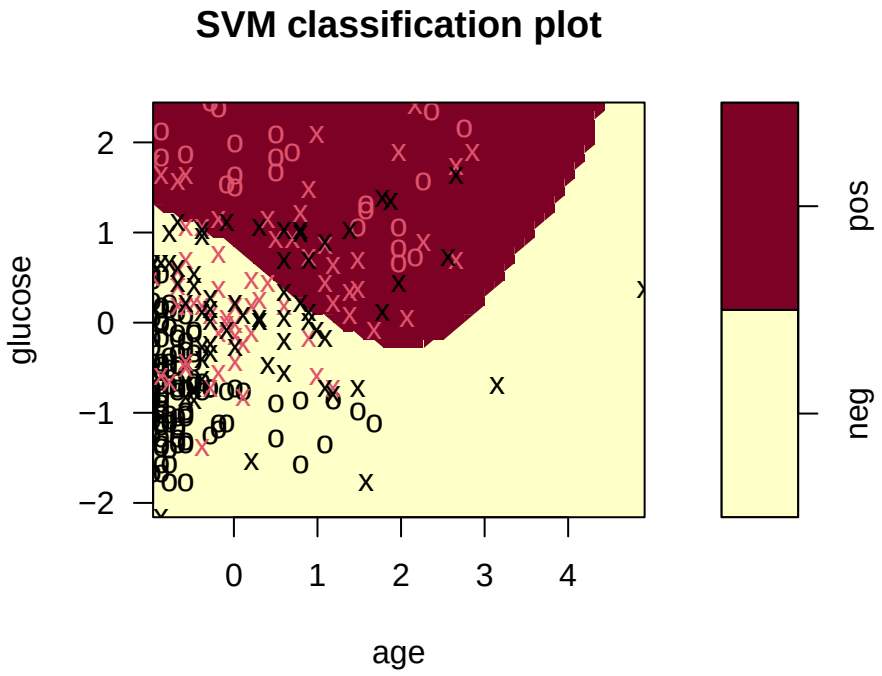


Rysunek 4: Wykres rozrzutu z podziałem na wynik klasyfikacji i rzeczywistą etykietkę dla danych uczących z wykorzystaniem wszystkich zmiennych, nbag=25, minsplit=0, cp=0

	parametry	error
1	train, nbag 25, ms 0, cp 0	0.000
2	train, nbag 25, ms 0, cp 0.02	0.000
3	train, nbag 25, ms 0, cp 0, g + a	0.008
4	train, nbag 25, ms 0, cp 0.02, g + a	0.008
5	test, nbag 25, ms 0, cp 0	0.218
6	test, nbag 25, ms 0, cp 0.02	0.233
7	test, nbag 25, ms 0, cp 0, g + a	0.286
8	test, nbag 25, ms 0, cp 0.02, g + a	0.286
9	crossvalidation, nbag 25, ms 0, cp 0	0.209
10	crossvalidation, nbag 25, ms 0, cp 0, g + a	0.250

Tabela 2: Zestawienie błądów klasyfikacji

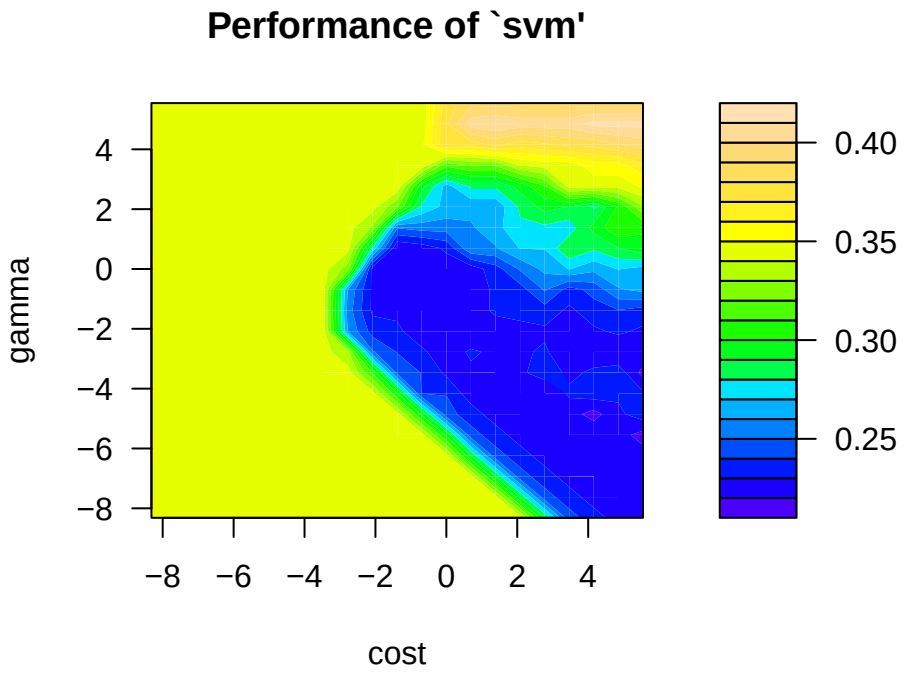
2.3 SVM



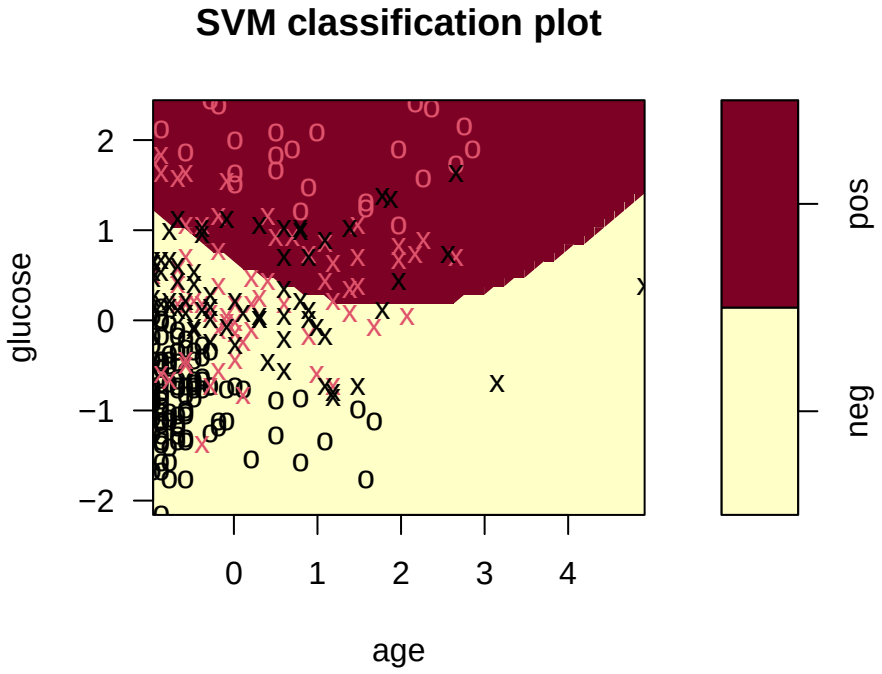
Rysunek 5: Klasyfikacja metodą wektorów nośnych z jądrem radialnym. gamma = 1, cost = 1, dane uczące

Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation
- best parameters: cost gamma 256 0.00390625
- best performance: 0.2161538



Rysunek 6: Wykres (mapa cieplna) skuteczności SVM w zależności od kosztu i gammy



Rysunek 7: Klasyfikacja SVM z wykorzystaniem optymalnych parametrów

	parametry	error
1	train, linear, cost 1	0.224
2	train, polynomial, degree 3, cost 1	0.243
3	train, radial, cost 1, gamma 0.5	0.212
4	train, linear, cost 0.5	0.224
5	train, polynomial, degree 10, cost 0.5	0.243
6	train, radial, cost 0.5, gamma 0.5	0.208
7	test, linear, cost 1	0.218
8	test, polynomial, degree 3, cost 1	0.226
9	test, radial, cost 1, gamma 0.5	0.195
10	test, linear, cost 0.5	0.218
11	test, polynomial, degree 10, cost 0.5	0.233
12	test, radial, cost 0.5, gamma 0.5	0.203
13	train, radial, cost tuned (256), gamma tuned (2^-10)	0.216
14	test, radial, cost tuned (256), gamma tuned (2^-10)	0.203

Tabela 3: Zestawienie błądów klasyfikacji

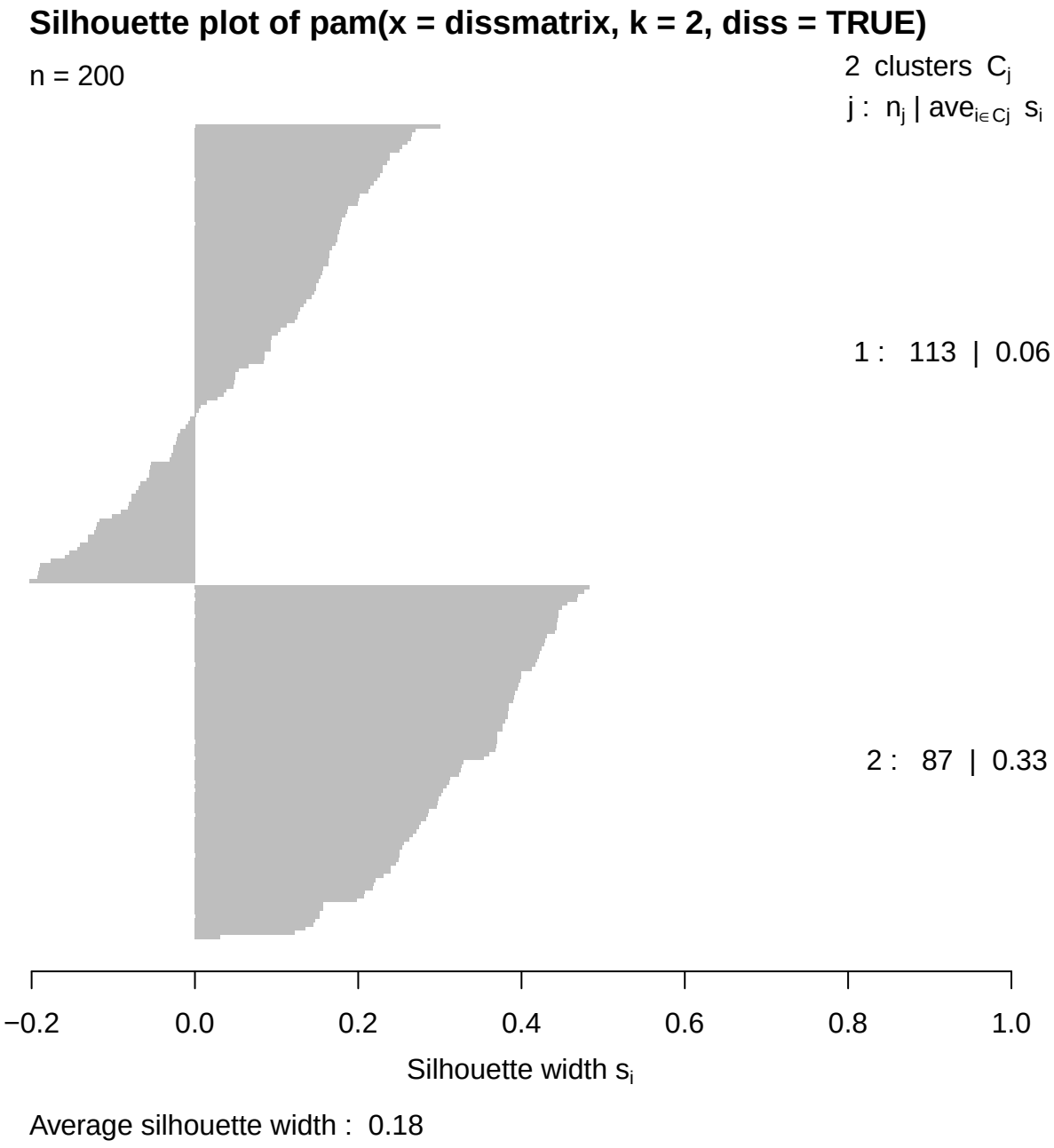
"Strojenie" parametrów nie przyniosło oszałamiających efektów, ale trzeba mieć na uwadze, że i bez strojenia skuteczność SVM była bardzo dobra. Metoda zdaje się nie być bardzo wrażliwa na dobór odpowiednich parametrów - nawet domyślne wartości zapewniły satysfakcjonujące rezultaty.

2.4 Podsumowanie

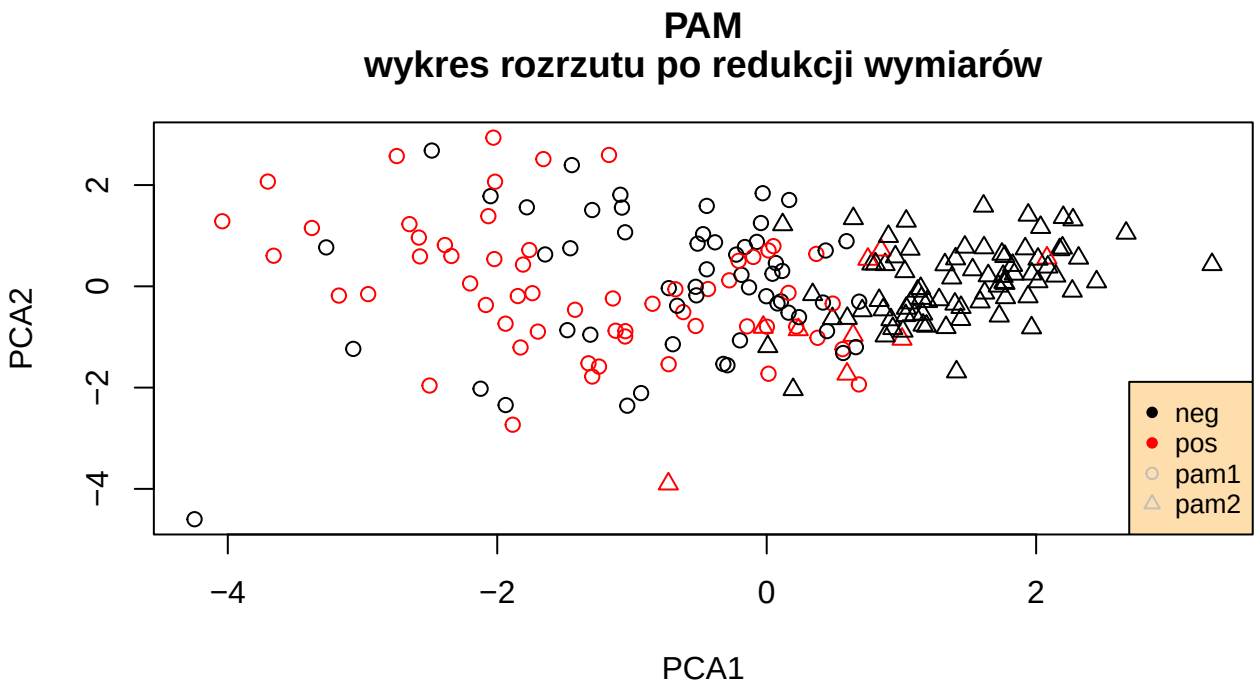
Każda z powyższych metod cechuje się wysoką dokładnością - wyższą, niż można oczekiwać większości standardowych metod stosowanych w klasyfikacji. Metoda SVM wyróżnia się przede wszystkim dobrymi efektami niezależnie od dobranych parametrów. Metoda boosting osiąga dobre wyniki dla większej ilości iteracji, jednak dłuższy czas działania jest odczuwalny. Metoda bagging działa dobrze, ale nie jest łatwo wybrać odpowiednie parametry. Biorąc pod uwagę powyższe, oceniam, że najlepszą skuteczność wykazała metoda wektorów nośnych.

3 Zadanie 2 - Analiza skupień – algorytmy grupujące i hierarchiczne

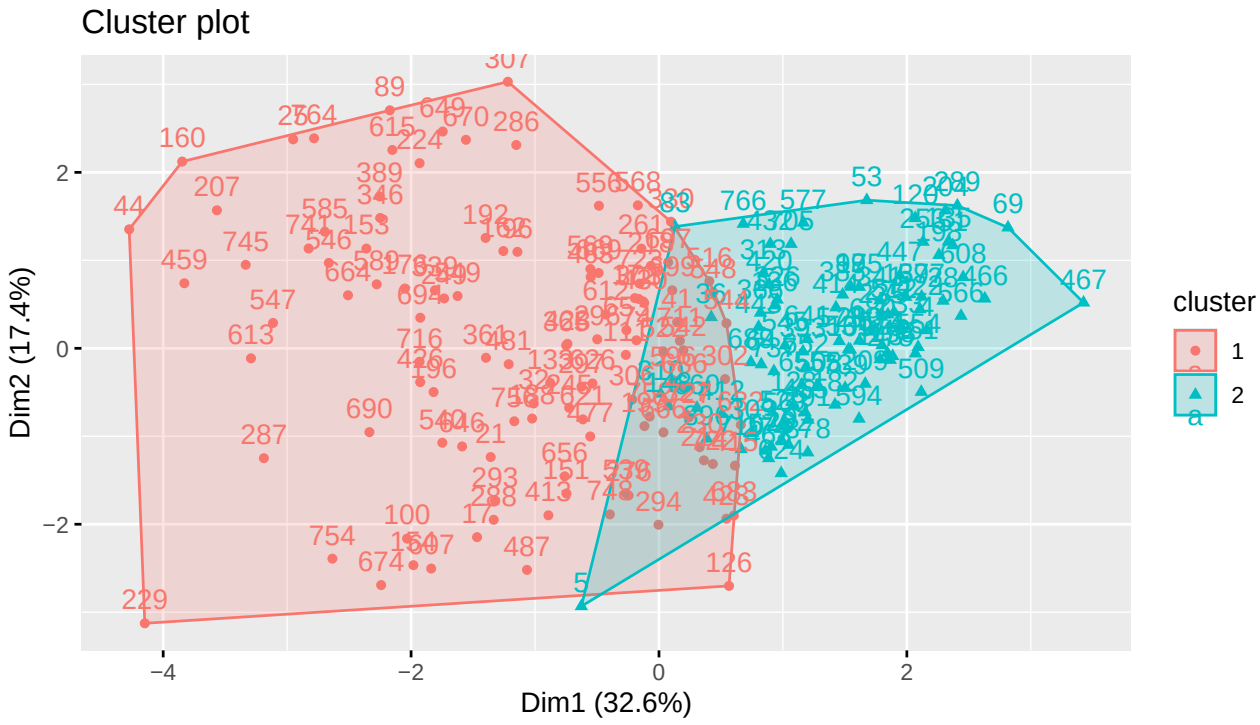
3.1 PAM (Partition around medoids)



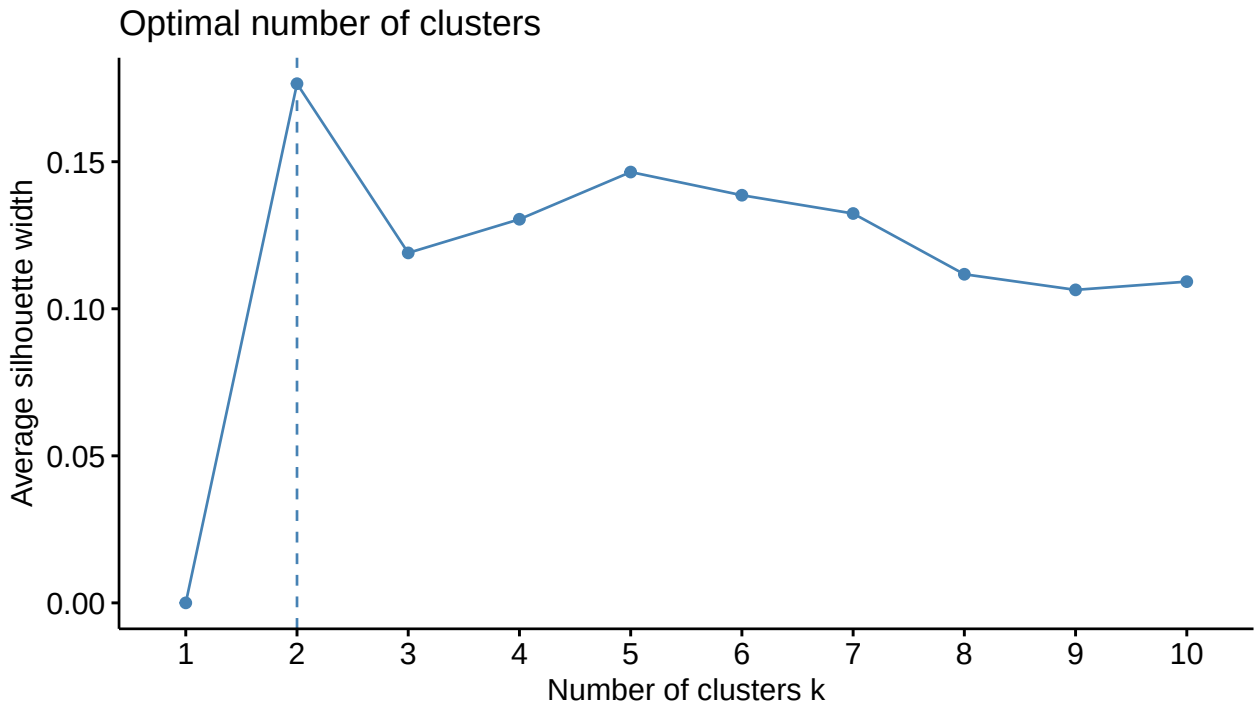
Rysunek 8: Wykres sillhouette - PAM



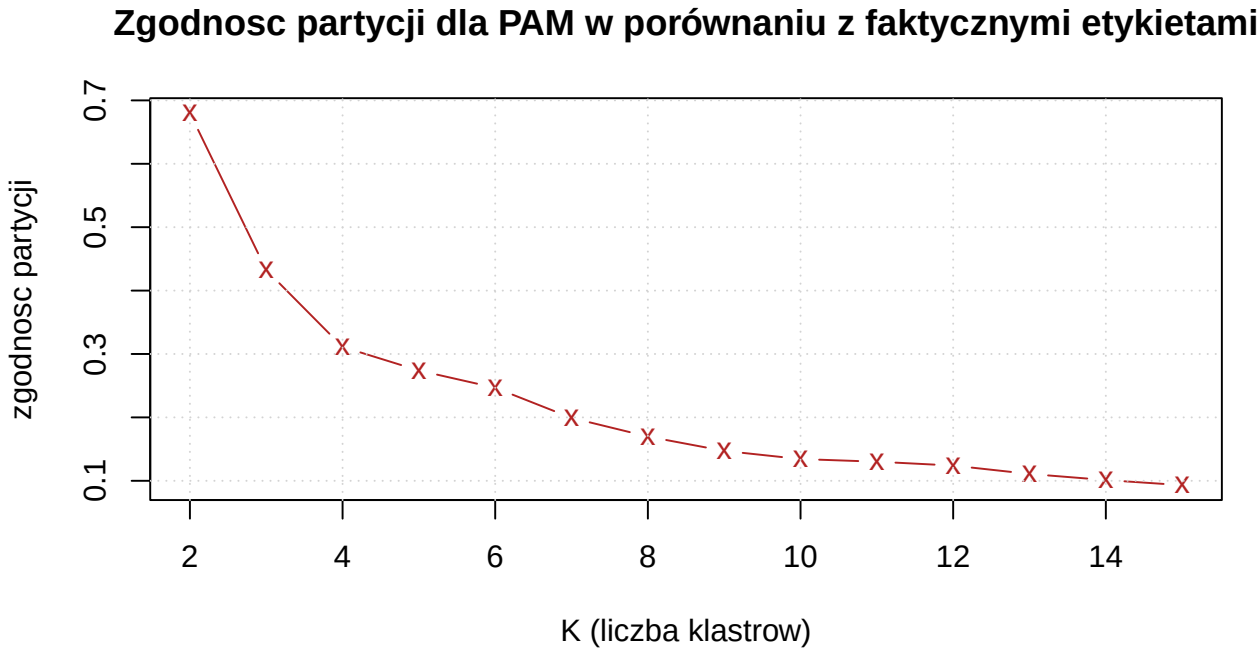
Rysunek 9: Wykres rozrzutu po redukcji wymiarów PCA z podziałem na wynik grupowania i rzeczywistą etykietkę



Rysunek 10: Wykres rozrzutu z zaznaczonymi klastrami



Rysunek 11: Wykres zależności średniej wartości silhouette od liczby klastrów

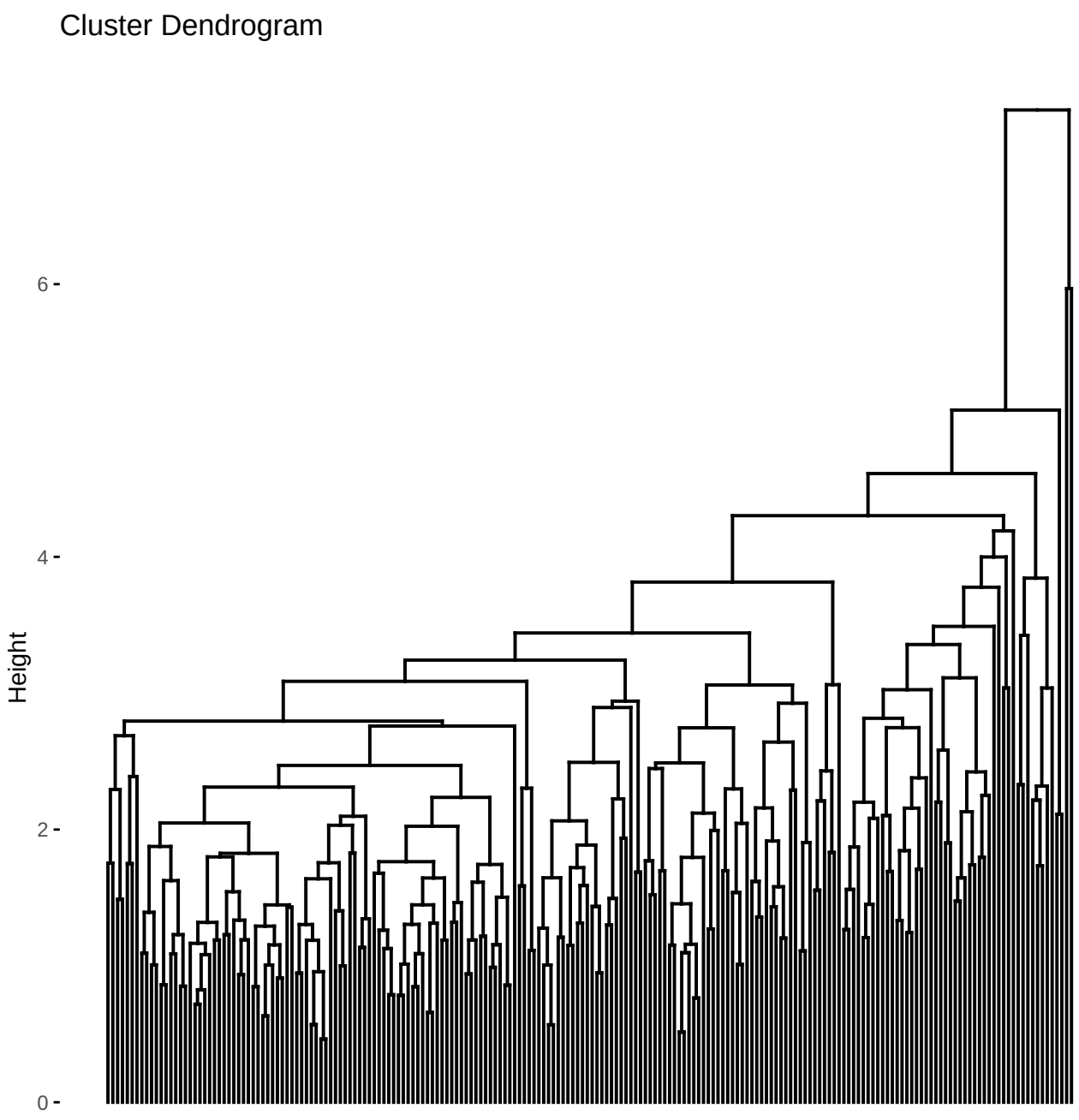


Rysunek 12: Wykres zależności zgodności partycji od liczby klastrów

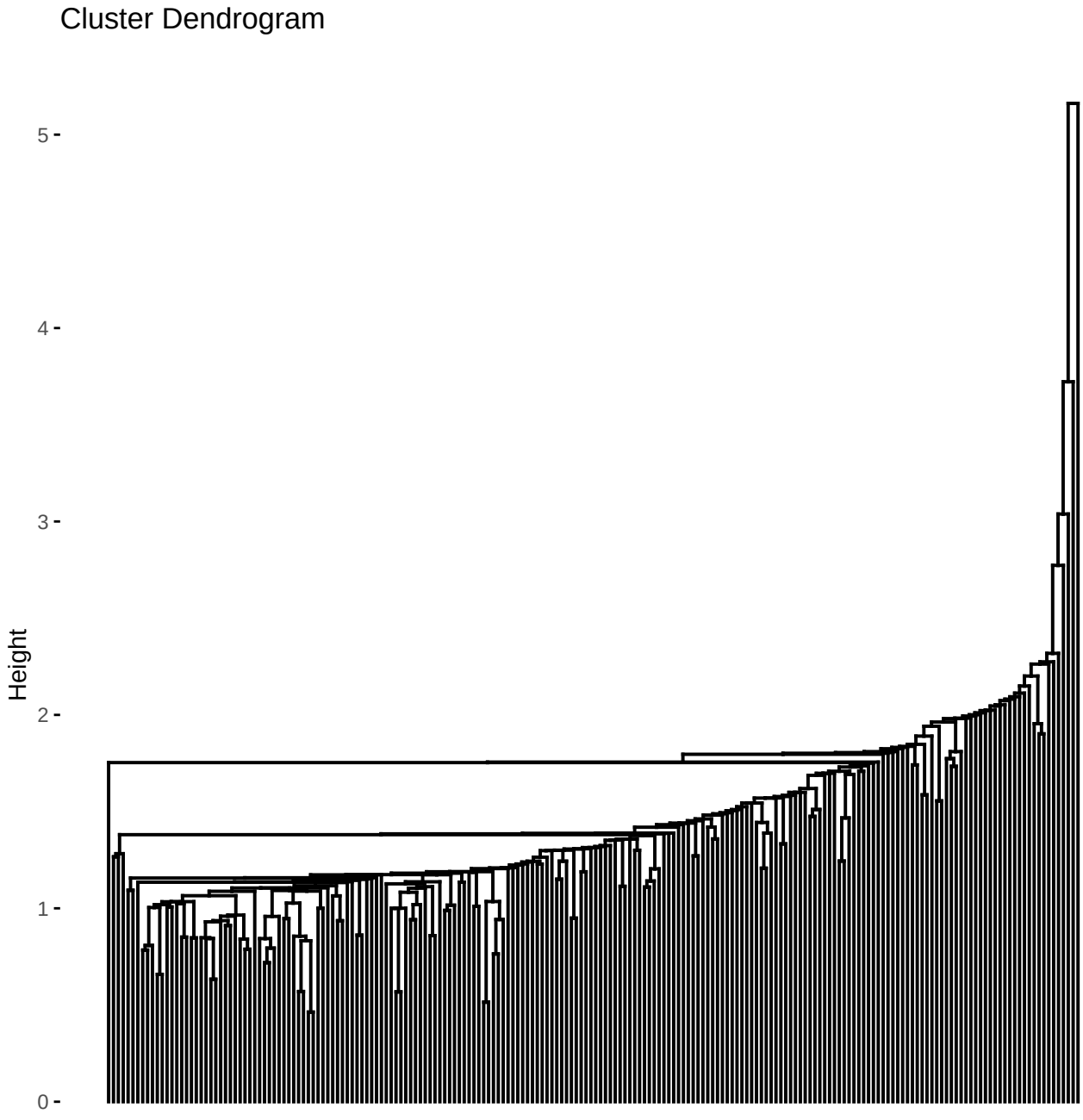
	1	2
neg	55	78
pos	58	9

Tabela 4: Macierz pomyłek dla metody PAM

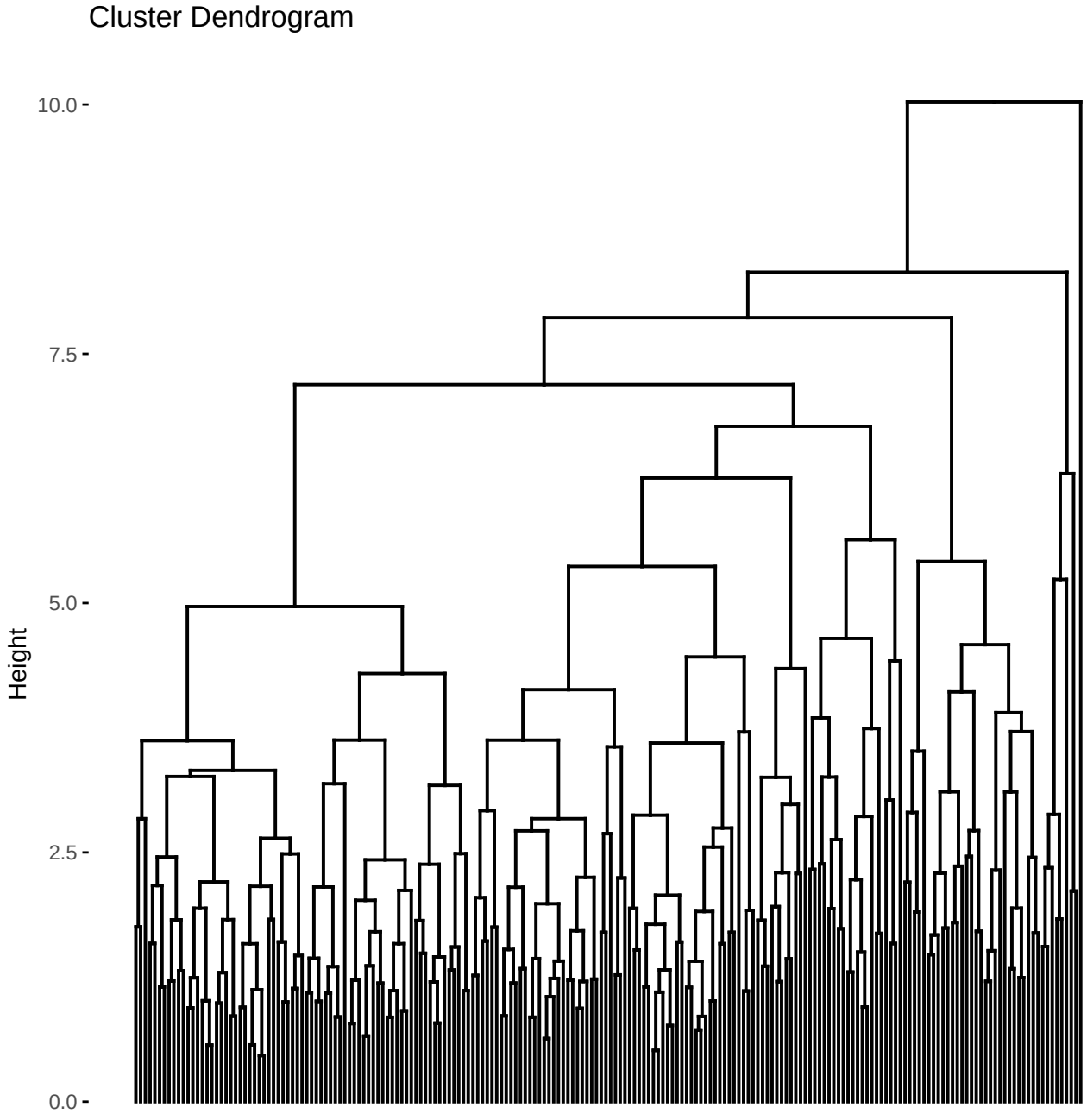
3.2 Agnes (Aglomerative nesting)



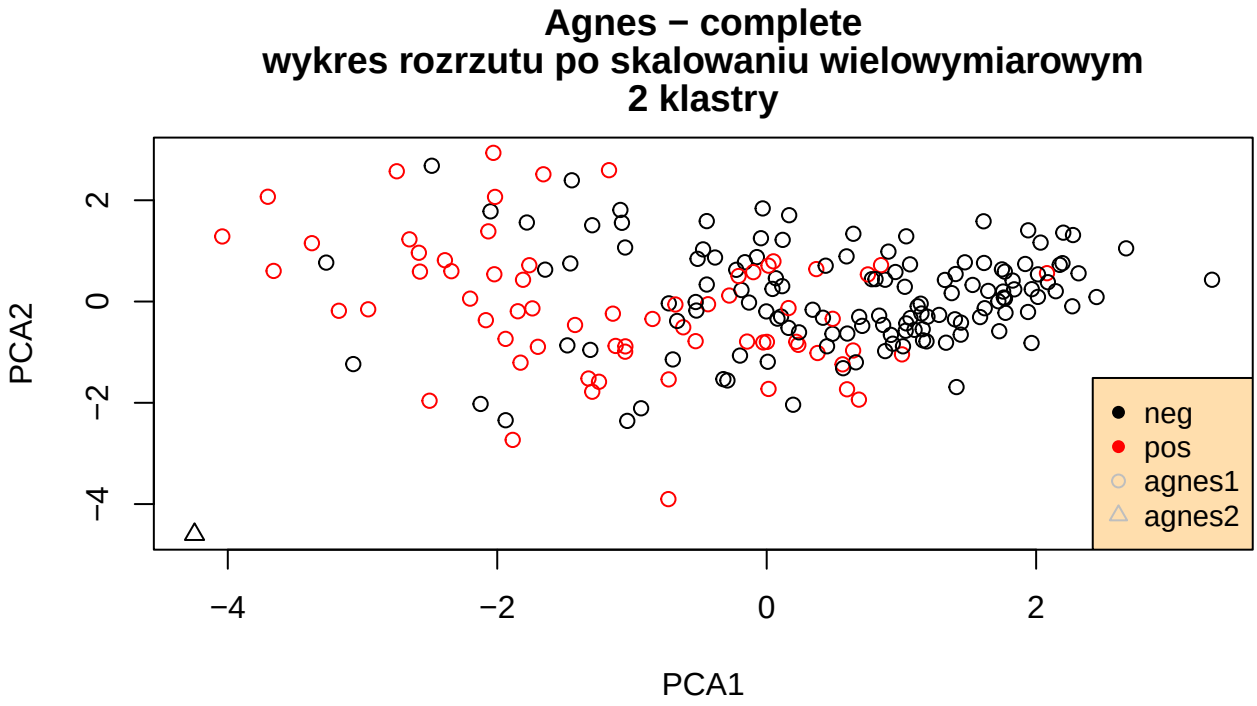
Rysunek 13: Dendrogram Agnes, average linkage.



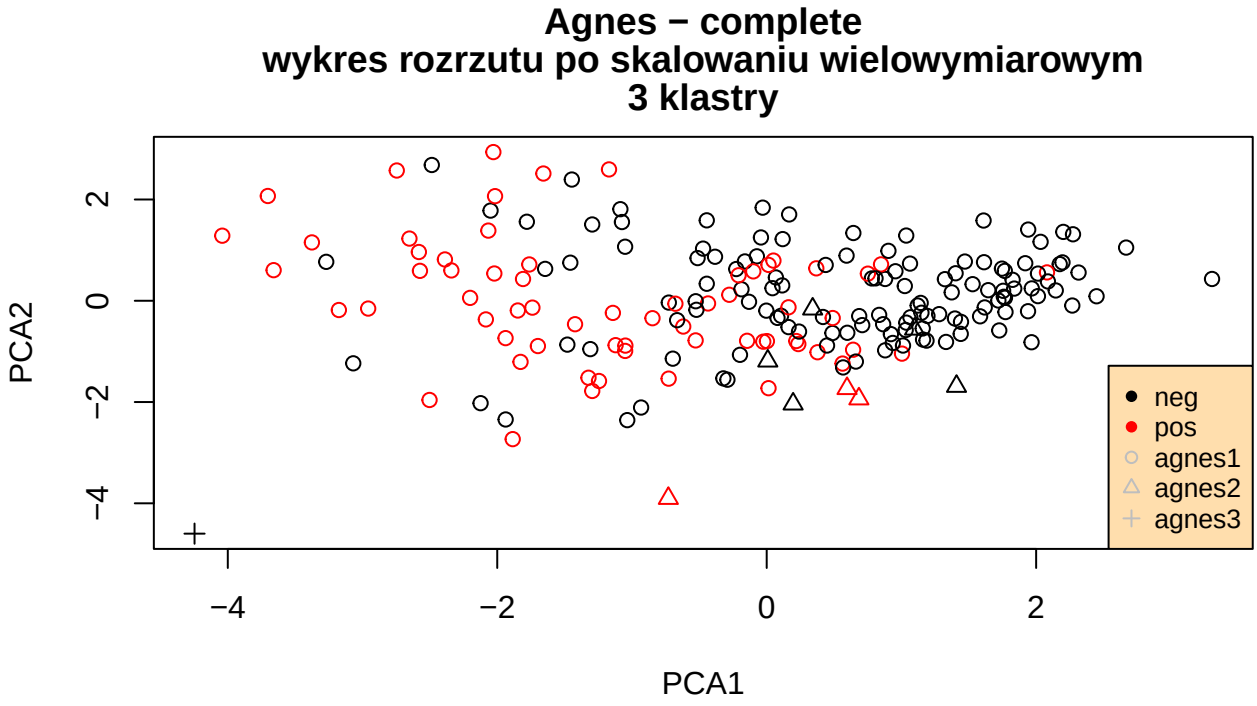
Rysunek 14: single linkage



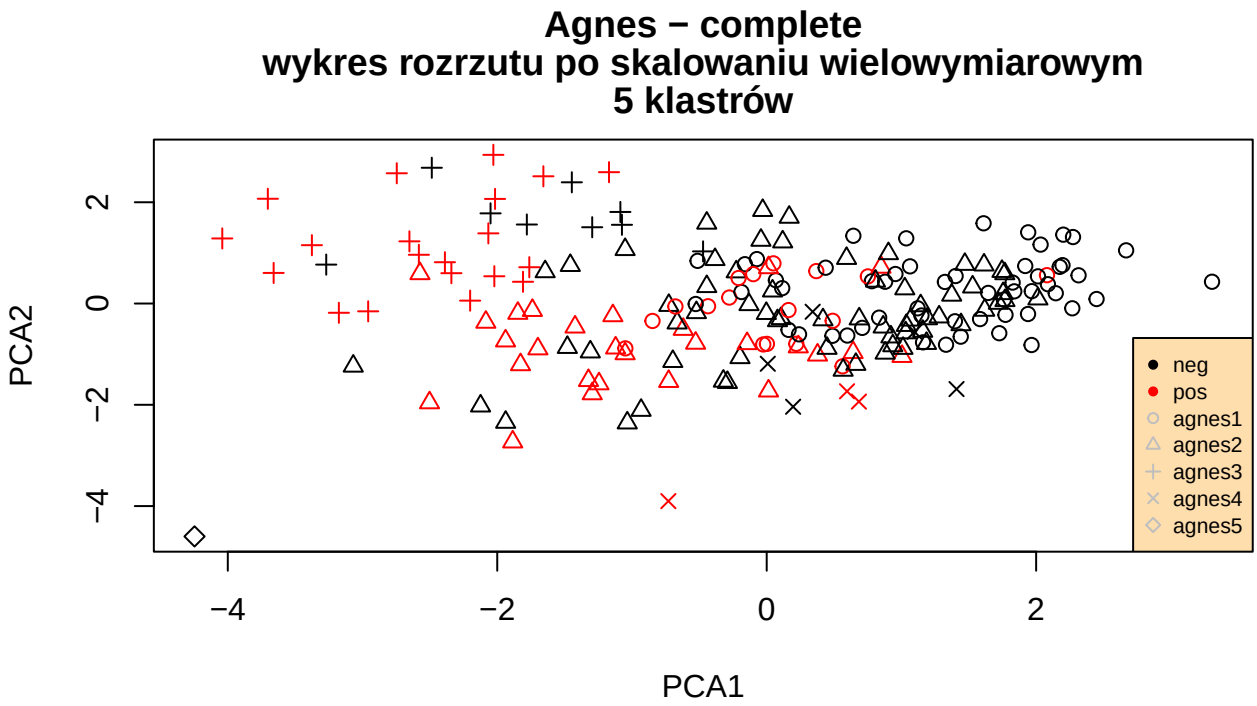
Rysunek 15: complete linkage



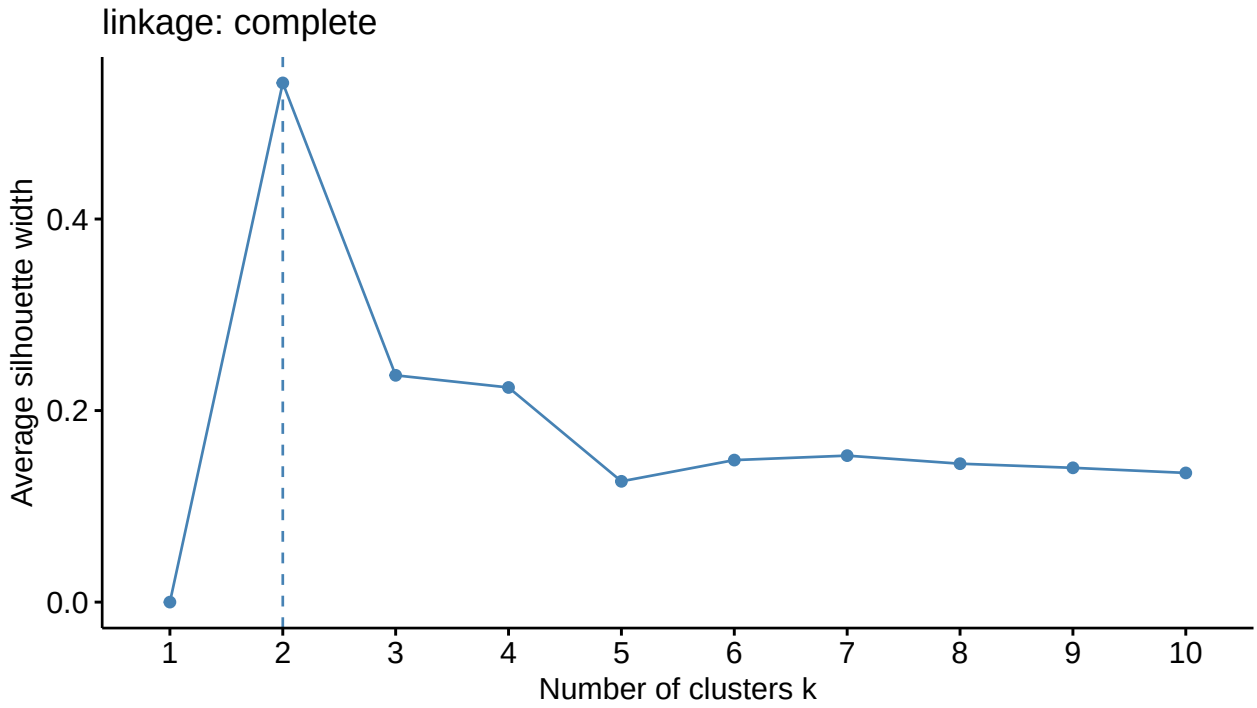
Rysunek 16: Wykres rozrzutu dla algorytmu Agnes z metodą najdalszego sąsiada. Liczba klastrów - 2



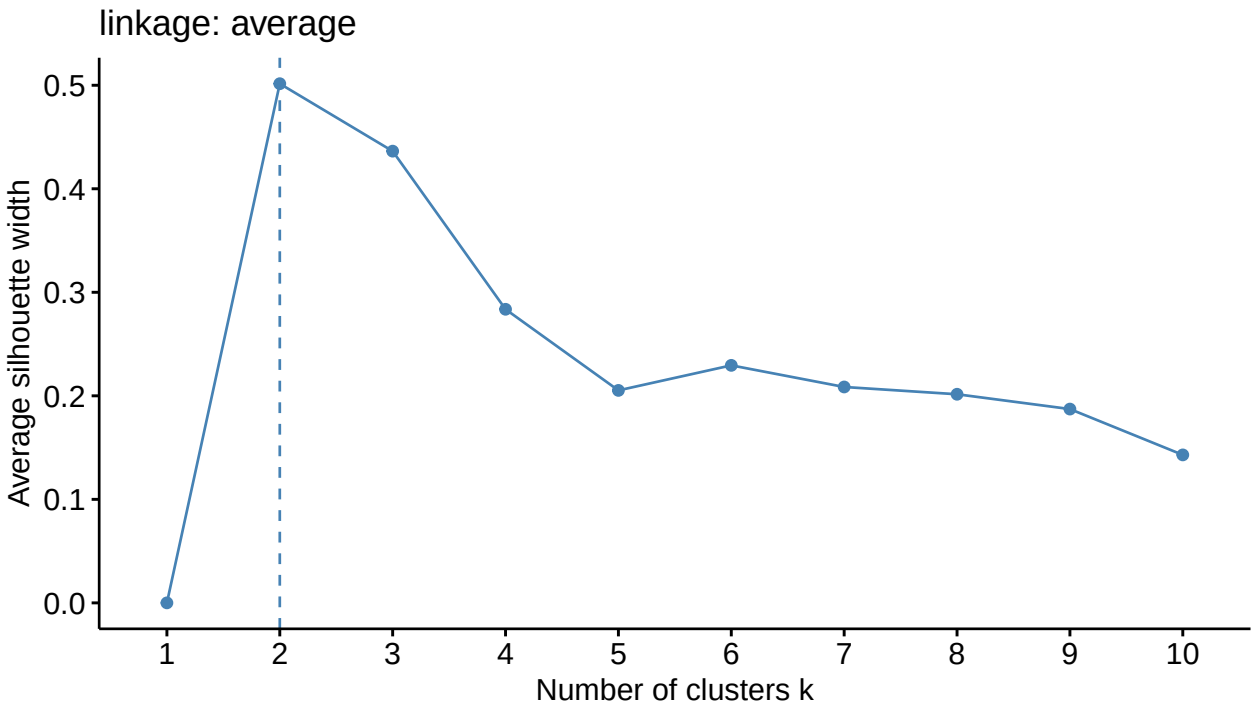
Rysunek 17: Wykres rozrzutu dla algorytmu Agnes z metodą najdalszego sąsiada. Liczba klastrów - 3



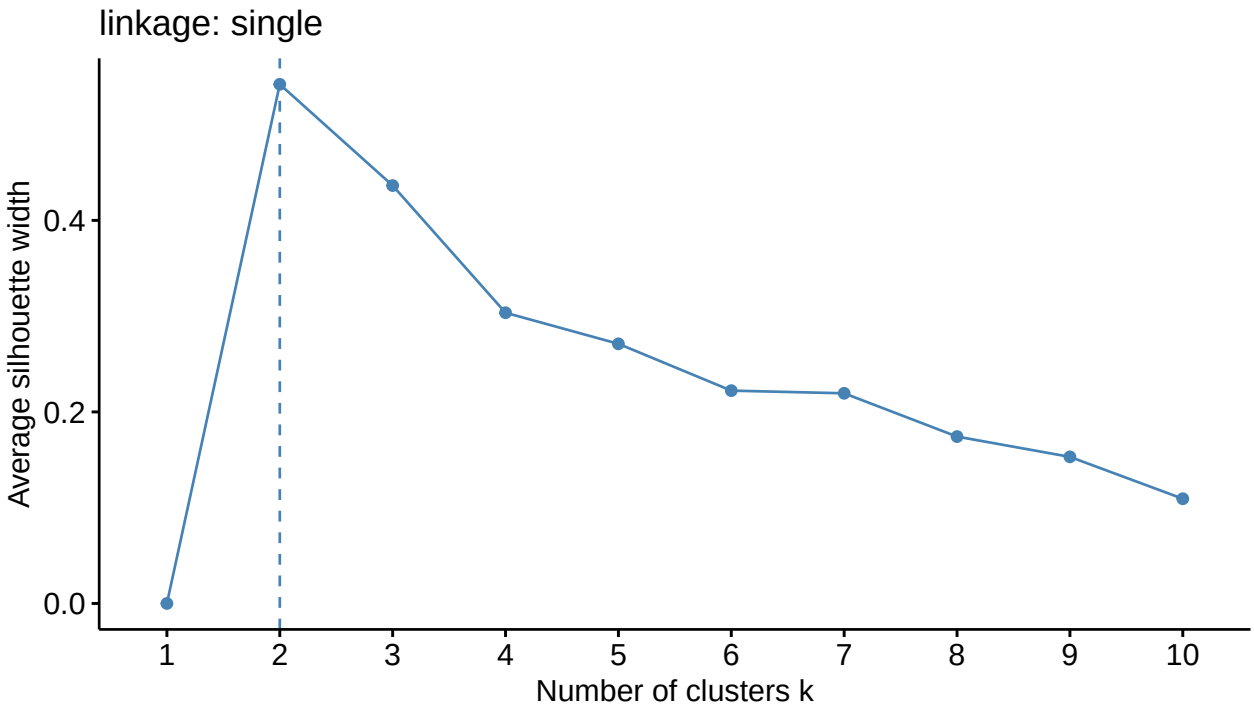
Rysunek 18: Wykres rozrzutu dla algorytmu Agnes z metodą najdalszego sąsiada. Liczba klastrów - 5



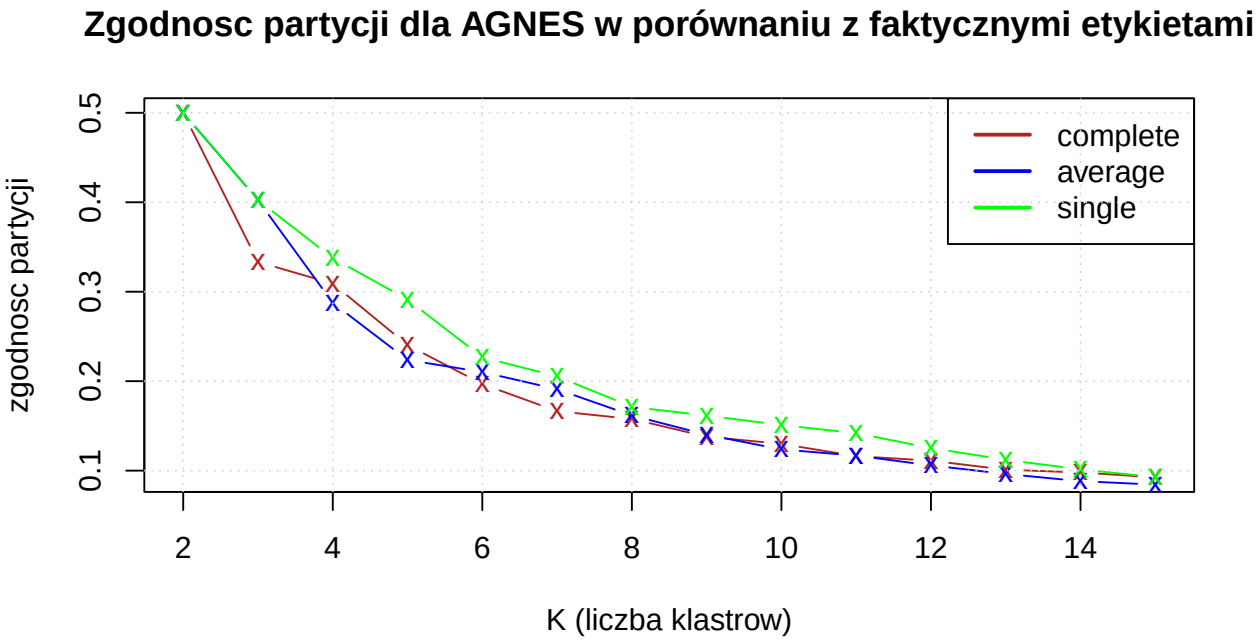
Rysunek 19: Wykres zależności średniej wartości silhouette od liczby klastrów dla algorytmu Agnes z metodą najdalszego sąsiada



Rysunek 20: Wykres zależności średniej wartości silhouette od liczby klastrów dla algorytmu Agnes z metodą odległości średniej

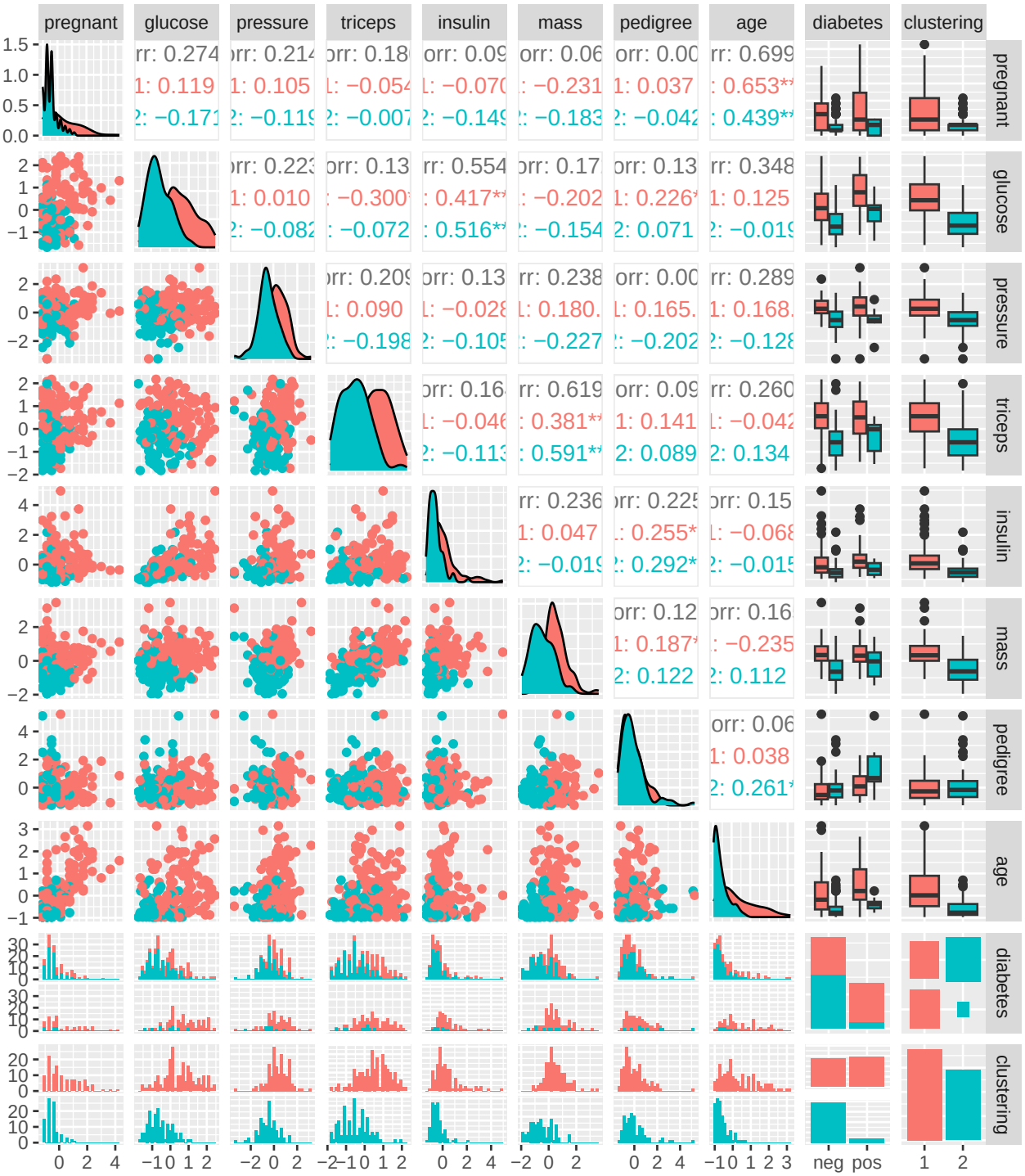


Rysunek 21: Wykres zależności średniej wartości silhouette od liczby klastrów dla algorytmu Agnes z metodą najbliższego sąsiada

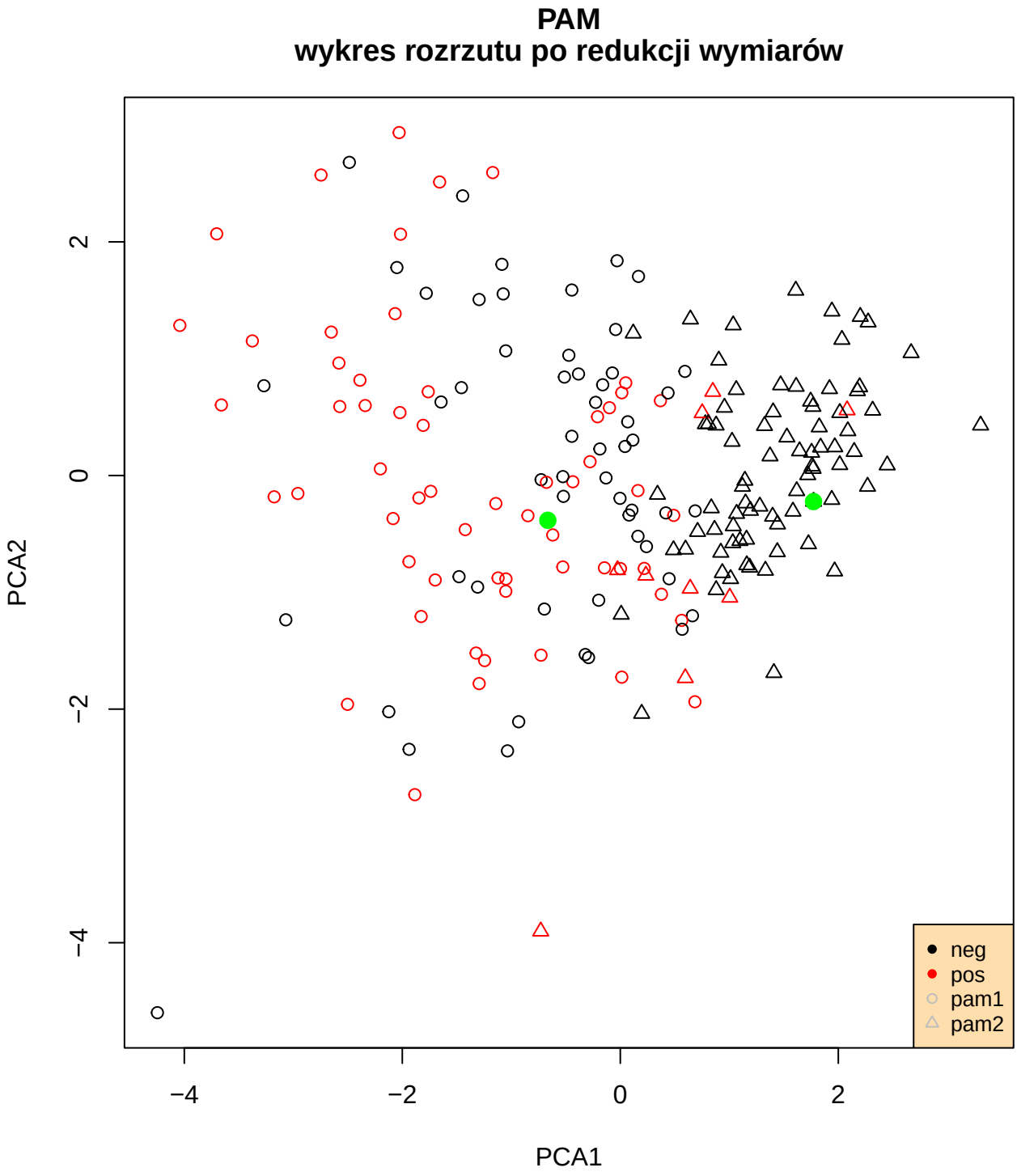


Rysunek 22: Wykres zgodności klastrów z faktycznymi etykietami (diabetes) z podziałem na metody łączenia klastrów

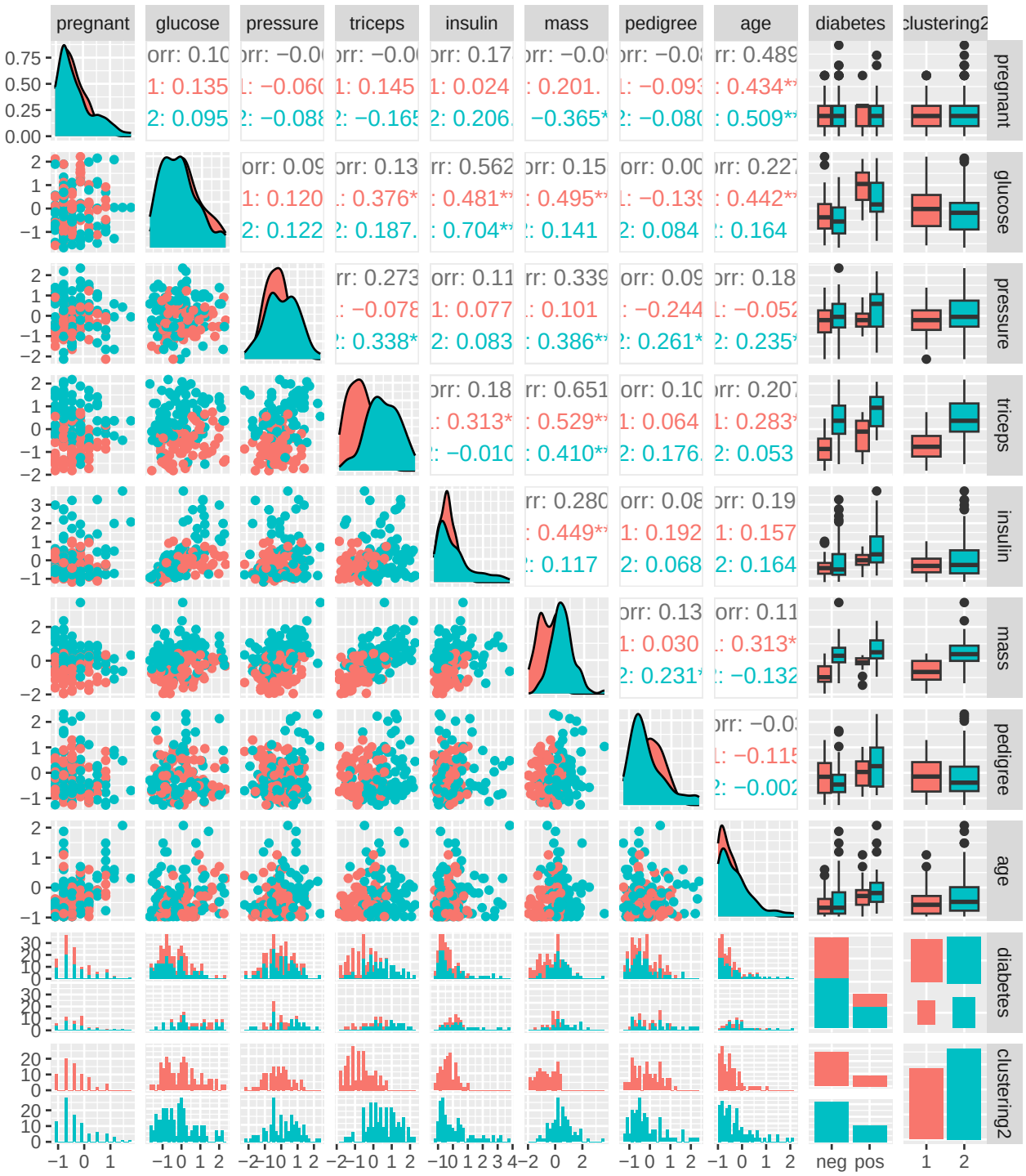
3.3 Podsumowanie



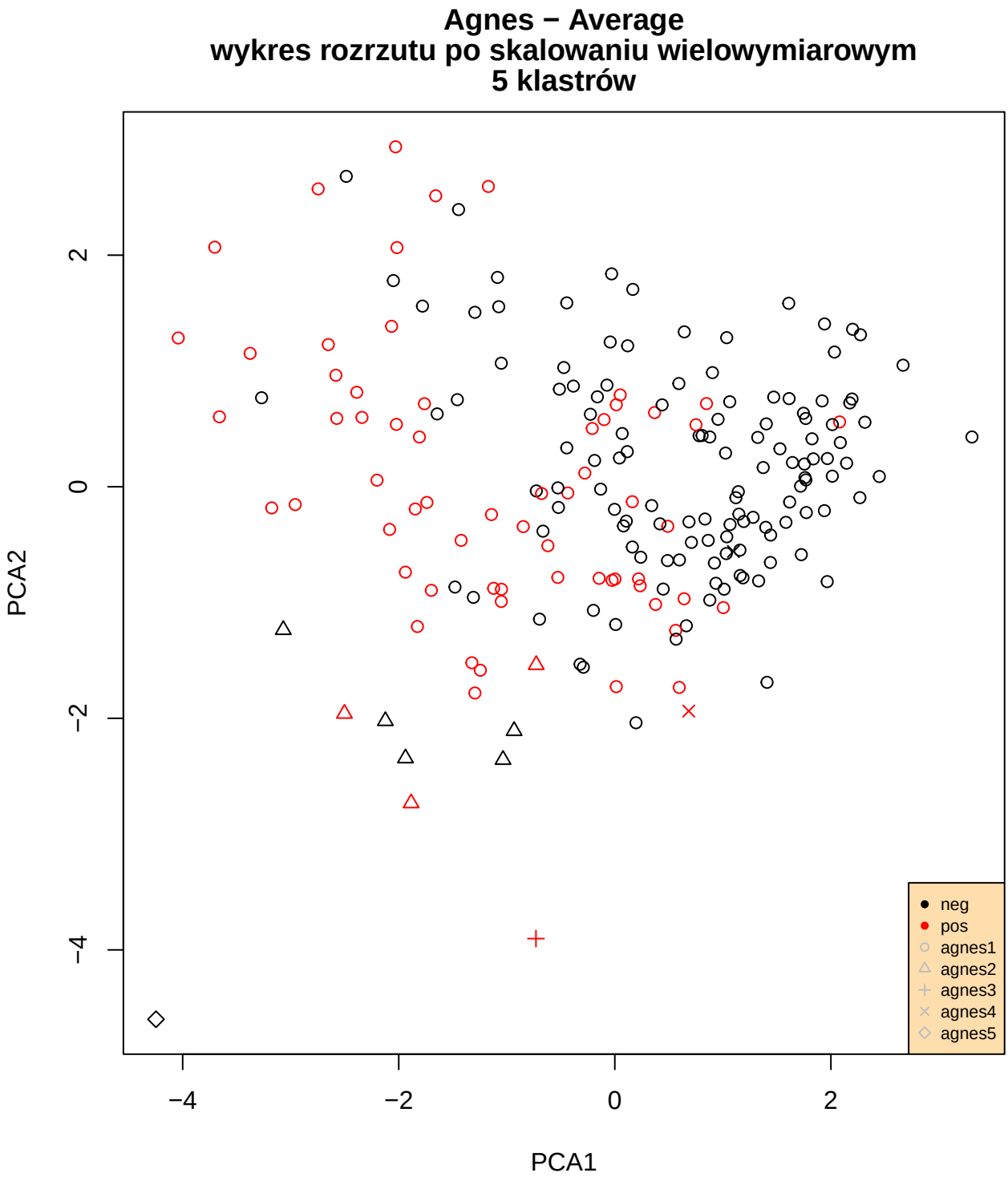
Rysunek 23: Zestawienie wykresów z podziałem na klastry algorytmu PAM.



Rysunek 24: Wykres rozrzutu (rys9) z zaznaczonymi medoidami



Rysunek 25: Zestawienie wykresów z podziałem na klastry algorytmu Agnes



Rysunek 26: Zestawienie wykresów z podziałem na klastry algorytmu PAM.

	pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
245	-0.405	0.757	0.427	0.557	0.319	0.728	-0.562	-0.183	neg
278	-1.028	-0.604	-0.533	-0.584	-0.337	-0.752	-0.200	-0.771	neg

Tabela 5: Cechy medoidów

Obie badane metody nie poradziły sobie dobrze jeśli chodzi o zestawienie z oryginalnymi etykietkami - ledwo (w najlepszym wypadku) udało im się osiągnąć poziom skuteczności związany z przydzieleniem wszystkich przypadków do większej kategorii. Klastry uzyskane za pomocą algorytmu agnes cechowały się wyższymi wartościami silhouette, co sugeruje lepszą zwartość i separację, lecz analiza wykresów składowych głównych podsuwa odwrotne wnioski. W przypadku algorytmu agnes dopiero przy ok. pięciu klastrach pojawiały się liczniejsze grupy. W związku z tym - i wcześniejszymi wykresami - oceniam, że optymalną liczbą klastrów dla agnes jest 5, a metodą - odległość średnia. Algorytm PAM daje najlepsze efekty zarówno pod kątem silhouette, jak i zgodności dla dwóch klastrów. Medoidy zdają się znajdować w centrum klastrów, jednak ciężko dostrzec coś szczególnego, jeśli chodzi o wartości ich cech.