

Indukcyjne metody analizy danych

Ćwiczenie 3

Wybrane metody klasteryzacji w oparciu o system R

Prowadzący: dr inż. Paweł Myszkowski

Student: Piotr Bielak, 218137

WT 17:05

Wrocław, 8 maja 2018r.

Spis treści

1 Wprowadzenie	3
1.1 Cel ćwiczenia	3
1.2 Algorytm K-Means	3
1.3 Algorytm PAM	3
1.4 Metryki	4
2 Wyniki eksperymentu	5
2.1 Zbiór "Diabetes"	5
2.2 Zbiór "Glass"	11
2.3 Zbiór "Wine"	17
2.4 Zbiór "Seeds"	23
2.5 Kroswalidacja Seeds	29

1 Wprowadzenie

1.1 Cel ćwiczenia

Celem ćwiczenia było zapoznanie się z algorytmami K-Means oraz PAM, które służą do grupowania (klasteryzacji) danych. Należało również zbadać i ocenić ich działanie na 4 określonych zbiorach danych (3 z poprzednich ćwiczeń oraz jeden wybrany zbiór, który jest typowy dla zagadnienia klasteryzacji). W trakcie badań należało uwzględnić różne parametry algorytmów, takie jak metryka odległości czy liczba klastrów, a następnie zaobserwować wpływ tych parametrów na wartości zadanych metryk.

1.2 Algorytm K-Means

Jest to jeden z najprostszych algorytmów klasteryzacji, należący do grupy algorytmów zachłannych (nie ma gwarancji znalezienia najlepszego rozwiązania). Parametrem jest liczba klastrów k . Opiera na się na idei wyliczenia k centroidów, dla każdego klastra po jednym. W tym celu algorytm próbuje zminimalizować tzw. błąd średniokwadratowy:

$$Err = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2,$$

gdzie:

k – liczba klastrów,

n – liczba obiektów,

$x_i^{(j)}$ – i -ty obiekt w j -tym klastrze,

c_j – centroid j -tego klastra,

$\|x_i^{(j)} - c_j\|^2$ – wybrana miara odległości między i -tym obiektem a centroidem.

Algorytm K-Means można przedstawić następująco:

1. Wybierz początkowe k centroidów (np. losowo).
2. Przyporządkuj każdy obiekt (instancję) do najbliższego centroida (*klasteryzacja*).
3. Wyznacz nowe pozycje centroidów.
4. Powtarzaj kroki 2, 3 do momentu aż centroidy nie będą zmieniać położenia (lub osiągnięcia innego warunku stopu).

1.3 Algorytm PAM

Algorytm ten należy do grupy *K-Medoids* i podobnie jak *K-Means* jest algorytmem zachłannym. Parametrem tutaj jest również liczba klastrów k . Zamiast wyliczać pozycje centroidów, wyznaczane są pozycje medoidów. Lista kroków tego algorytmu jest następująca:

1. Wybierz początkowe k medoidów (np. losowe obiekty / instancje).
2. Przyporządkuj każdy obiekt do najbliższego medoida.
3. Dopóki można ulepszyć obecne rozwiązanie, dla każdego medoida m , dla każdego nie-medoida o wykonuj:
 - Zamień m oraz o i przelicz koszt (suma odległości obiektów od medoidów).
 - Jeśli całkowity koszt wzrosł, odrzuć zamianę.

Algorytm PAM można również przedstawić w oparciu o dwie fazy: *BUILD* (wybór początkowego zbioru medoidów) oraz *SWAP* (zamiana par m oraz o , takich aby jak najbardziej polepszyć klasteryzację). Dodatkowo zamiast obliczać bezpośrednio odległość między obiektemi, stosuje się tutaj miary **niepodobieństwa** (ang. *dissimilarity*) między danym obiektem a najbliższym oraz drugim najbliższym medoidem.

1.4 Metryki

W celu oceny jakości klasteryzacji zbiorów danych, użyto następujących miar / metryk:

- Davies-Bouldin Index (DBI) – miara wew.; bierze pod uwagę rozrzut instancji wewnętrz klastra oraz odległości między klastrami; wartość tej miary powinna być minimalizowana (lepsze są klastry o mały rozrzucie i odległe od siebie); jest zdefiniowana następująco:

$$DBI = \frac{1}{K} \sum_{k=1}^K M_k = \frac{1}{K} \sum_{k=1}^K \max_{k' \neq k} \left(\frac{\delta_k + \delta_{k'}}{\Delta_{kk'}} \right),$$

gdzie:

K – liczba klastrów,

δ_k – średnia odległość instancji w klastrze k od centroida,

$\Delta_{kk'}$ – odległość między centroidami klastrów k oraz k' .

- Dunn Index – miara wew.; bierze pod uwagę odległości między instancjami w różnych klastrach oraz tym samym klastrze; wartość tej miary powinna być maksymalizowana; jest zdefiniowana następująco:

$$Dunn = \frac{d_{min}}{d_{max}},$$

gdzie:

d_{min} – minimalna odległość między punktami należącymi do różnych klastrów (spośród wszystkich par klastrów),

d_{max} – maksymalna odległość między punktami w ramach jednego klastra (spośród wszystkich klastrów).

- Rand – miara zew.; dla każdej pary instancji sprawdzane jest czy zostały one przypisane do tego samego klastra; wymagane są tutaj dwie metody klasteryzacji / dwa wyniki klasteryzacji (można podać jako drugi wynik prawdziwe etykiety danych); jest zdefiniowana następująco:

$$Rand = \frac{a + b}{\binom{N}{2}},$$

gdzie:

a – liczba par instancji przypisanych do tego samego klastra,

b – liczba par instancji przypisanych do różnych klastrów,

N – liczba instancji.

- Purity – miara zew.; opiera się na liczbie wystąpień najliczniejszej klasy instancji w każdym z klastrów; jest zdefiniowana następująco:

$$Purity = \frac{1}{N} \sum_{k \in K} \max_{d \in D} |k \cap d|,$$

gdzie:

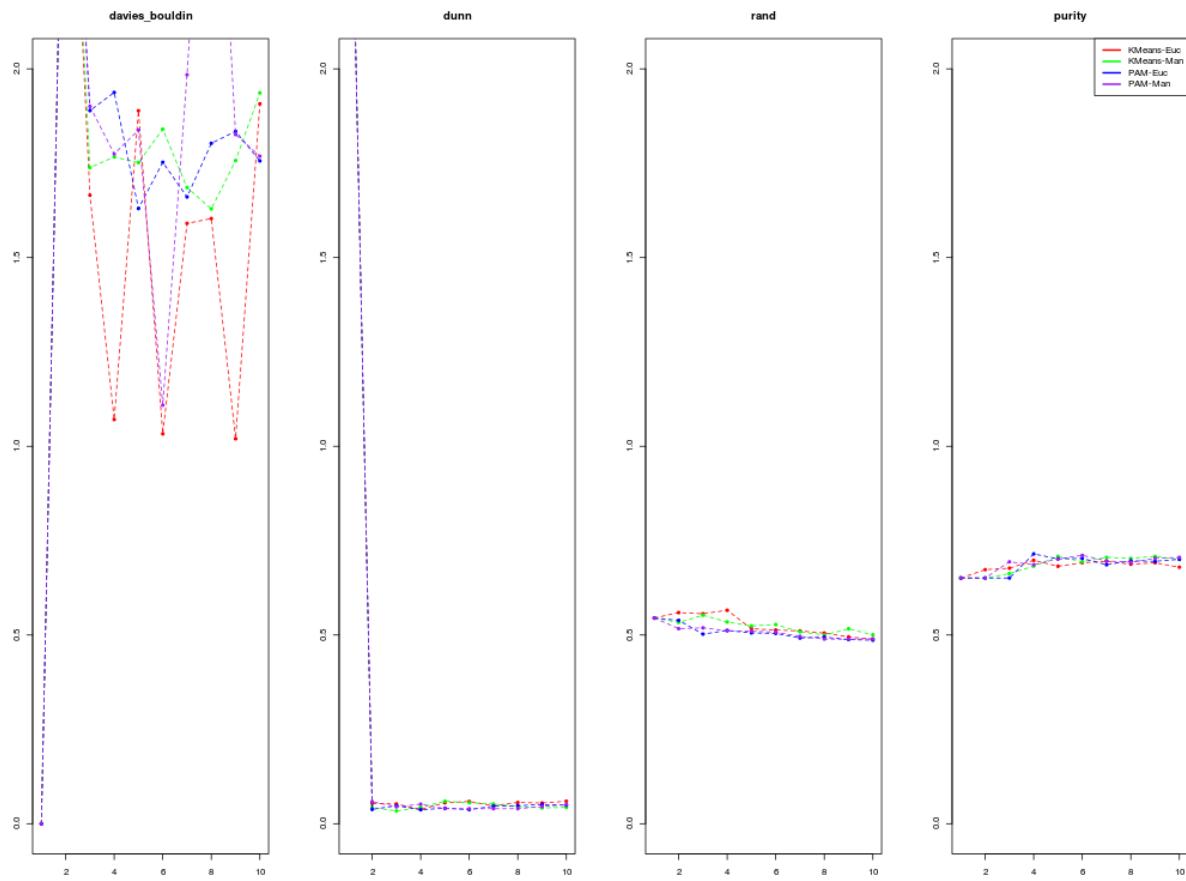
N – liczba instancji,

K – zbiór klastrów,

D – zbiór klas.

2 Wyniki eksperymentu

2.1 Zbiór "Diabetes"

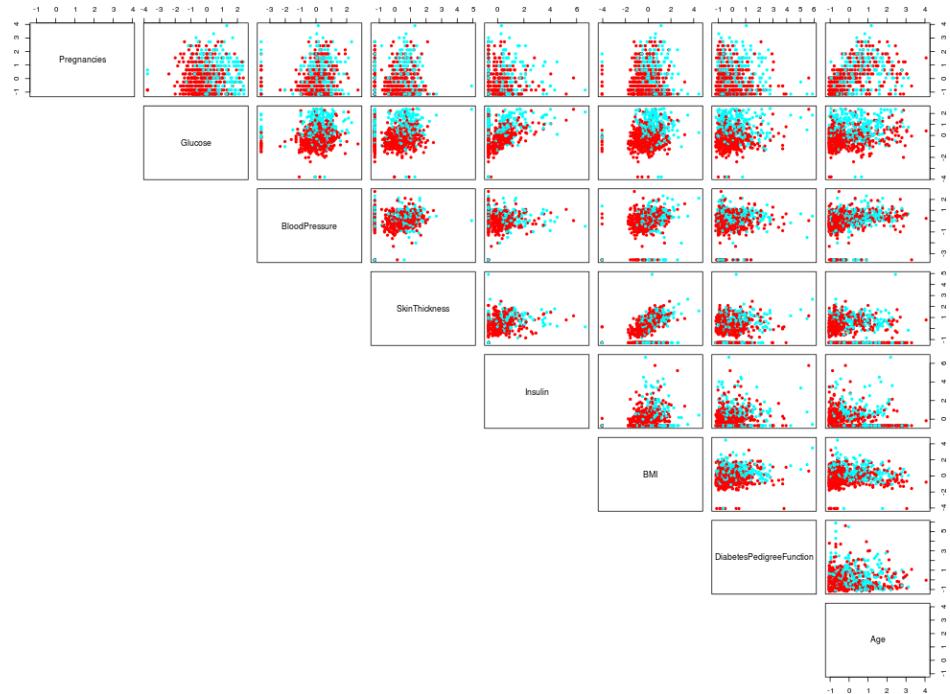


Rysunek 1: Wykresy wartości metryk dla zbioru "Diabetes".

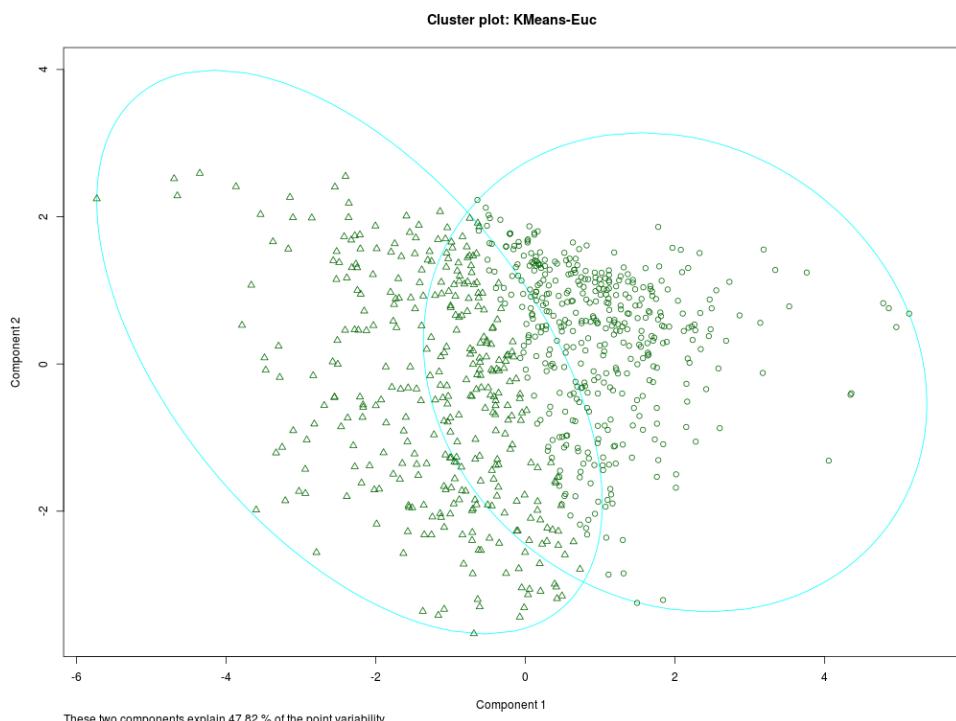
Algorytm	Liczba klastrów	davies_bouldin	dunn	rand	purity
KMeans-Euc	1	0.00	INF	0.55	0.65
	2	2.01	0.053	0.56	0.67
	3	1.67	0.052	0.56	0.68
	4	1.07	0.038	0.57	0.70
	5	1.89	0.056	0.52	0.68
	6	1.03	0.059	0.51	0.69
	7	1.59	0.048	0.51	0.69
	8	1.60	0.056	0.51	0.69
	9	1.02	0.055	0.49	0.69
	10	1.91	0.06	0.49	0.68
KMeans-Man	1	0.00	INF	0.55	0.65
	2	2.16	0.043	0.53	0.65
	3	1.74	0.034	0.55	0.66
	4	1.77	0.043	0.54	0.68
	5	1.75	0.06	0.52	0.71
	6	1.84	0.056	0.53	0.69
	7	1.69	0.052	0.51	0.71
	8	1.63	0.046	0.50	0.70
	9	1.76	0.042	0.52	0.71
	10	1.94	0.043	0.50	0.70
PAM-Euc	1	0.00	INF	0.55	0.65
	2	2.19	0.038	0.54	0.65
	3	1.89	0.048	0.50	0.65
	4	1.94	0.037	0.51	0.71
	5	1.63	0.041	0.51	0.70
	6	1.75	0.037	0.50	0.70
	7	1.66	0.047	0.49	0.69
	8	1.80	0.048	0.49	0.70
	9	1.83	0.051	0.49	0.69
	10	1.76	0.049	0.49	0.70
PAM-Man	1	0.00	INF	0.55	0.65
	2	2.35	0.058	0.52	0.65
	3	1.90	0.045	0.52	0.69
	4	1.77	0.052	0.51	0.69
	5	1.84	0.04	0.51	0.70
	6	1.11	0.04	0.51	0.71
	7	1.98	0.04	0.50	0.70
	8	2.02	0.04	0.49	0.69
	9	1.83	0.046	0.49	0.70
	10	1.77	0.051	0.49	0.71

Tabela 1: Wartości metryk dla zbioru "Diabetes".

2.1.1 Algorytm K-Means (Euclidean)

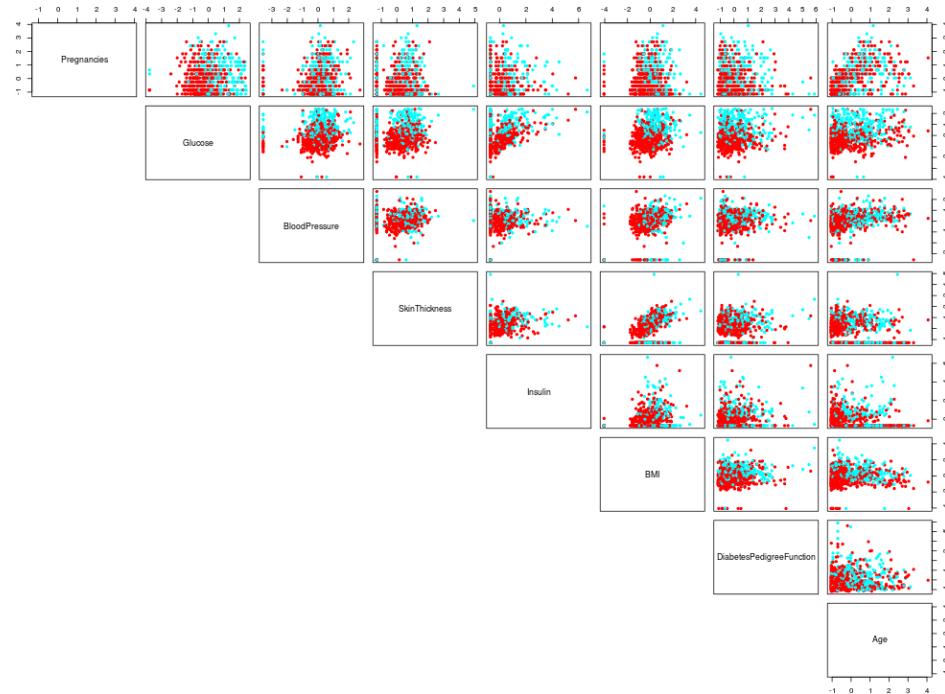


Rysunek 2: Wynik klasteryzacji dla algorytmu KMeans (Euclidean) dla zbioru "Diabetes".

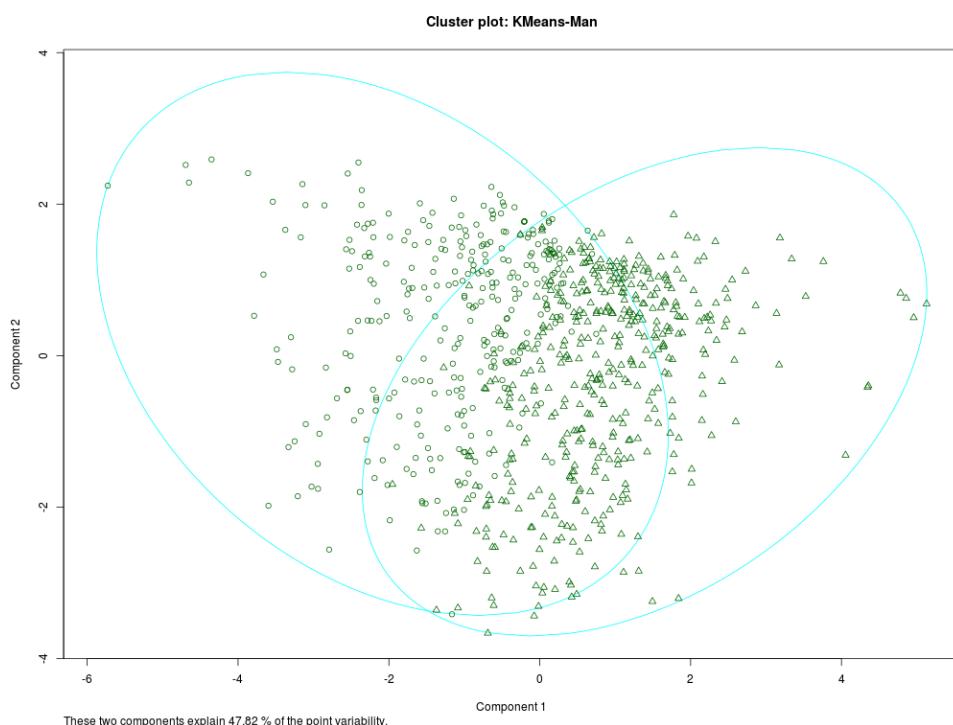


Rysunek 3: Klastry dla algorytmu KMeans (Euclidean) dla zbioru "Diabetes".

2.1.2 Algorytm K-Means (Manhattan)

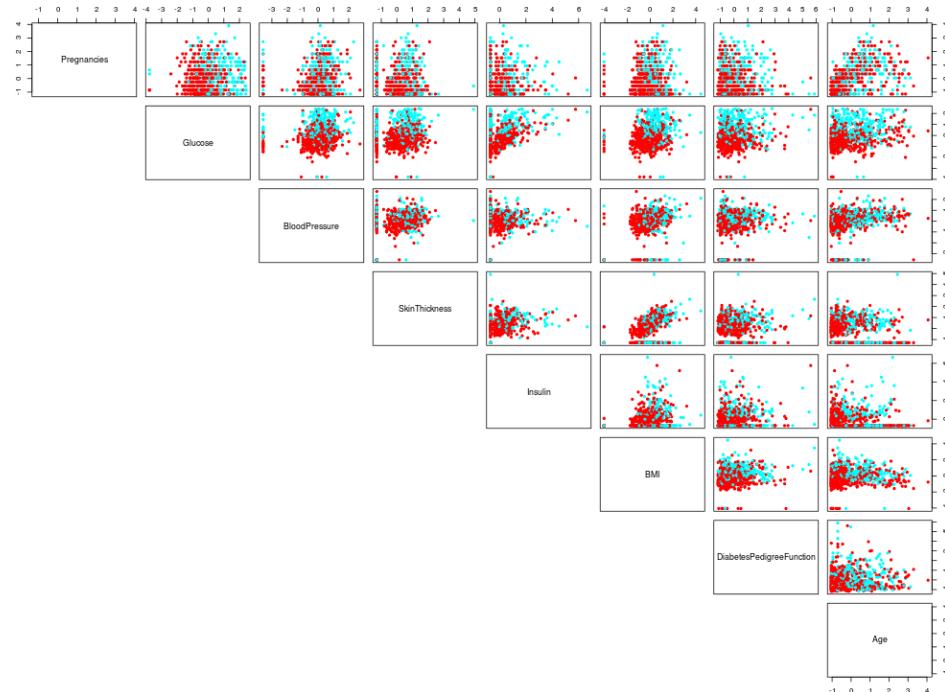


Rysunek 4: Wynik klasteryzacji dla algorytmu KMeans (Manhattan) dla zbioru "Diabetes".

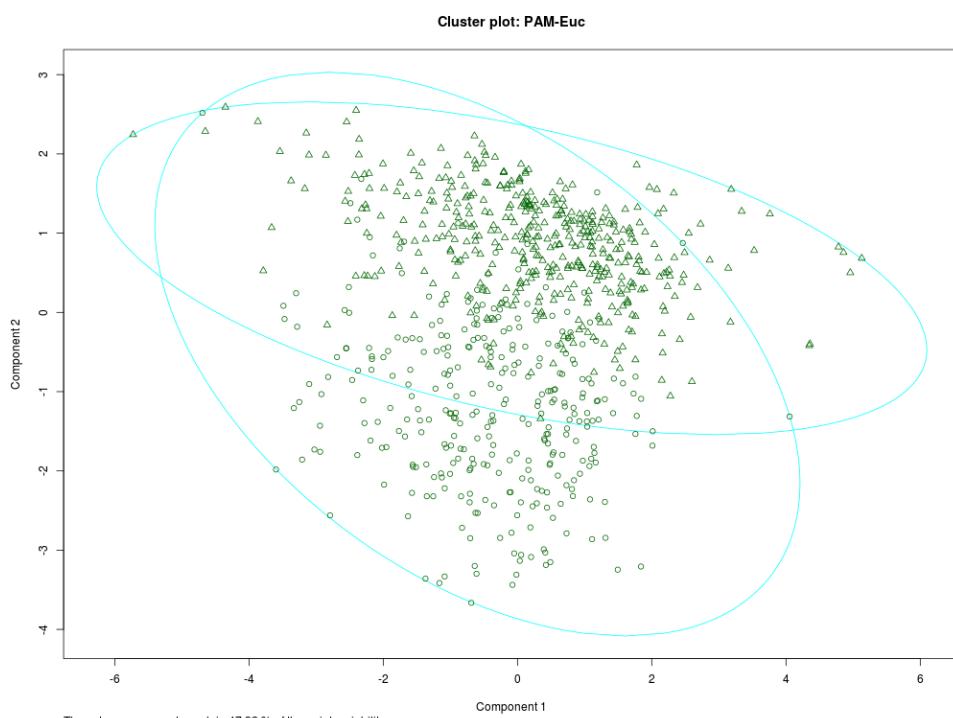


Rysunek 5: Klastry dla algorytmu KMeans (Manhattan) dla zbioru "Diabetes".

2.1.3 Algorytm PAM (Euclidean)

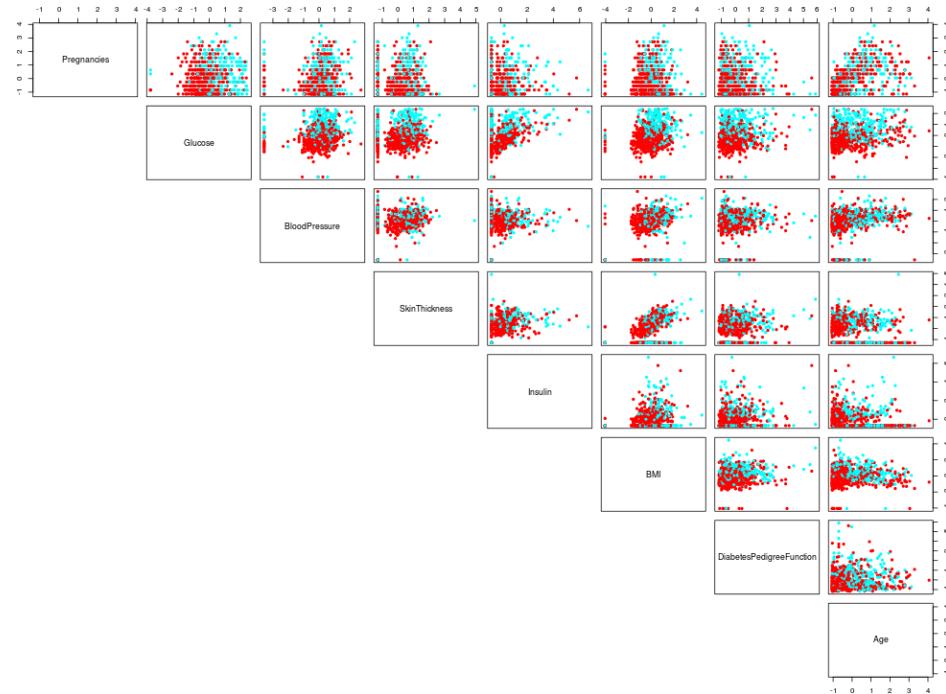


Rysunek 6: Wynik klasteryzacji dla algorytmu PAM (Euclidean) dla zbioru "Diabetes".

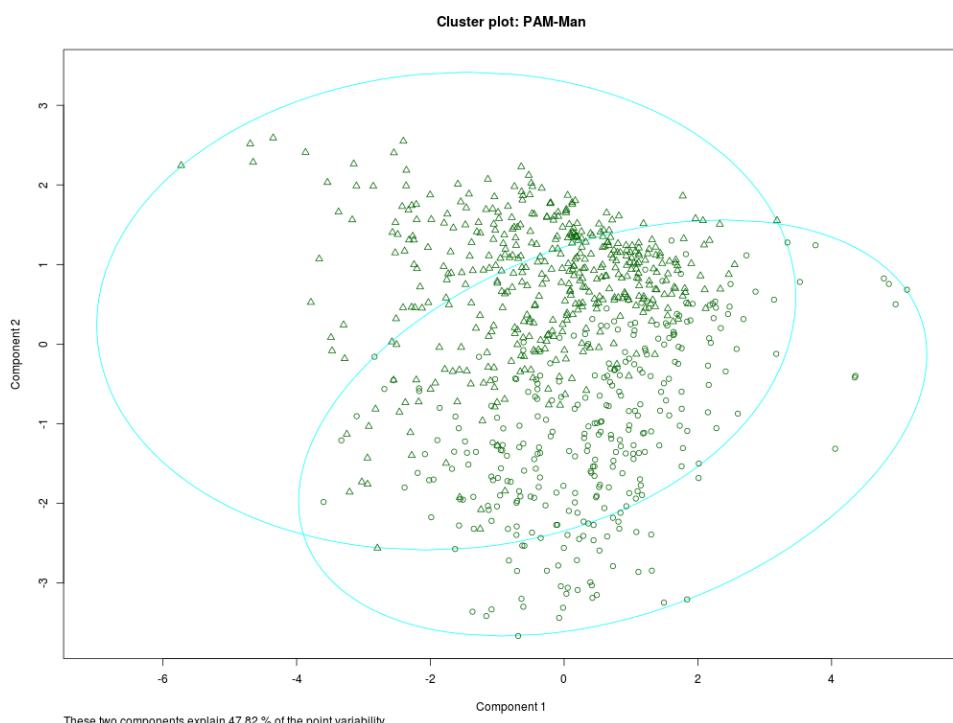


Rysunek 7: Klastry dla algorytmu PAM (Euclidean) dla zbioru "Diabetes".

2.1.4 Algorytm PAM (Manhattan)

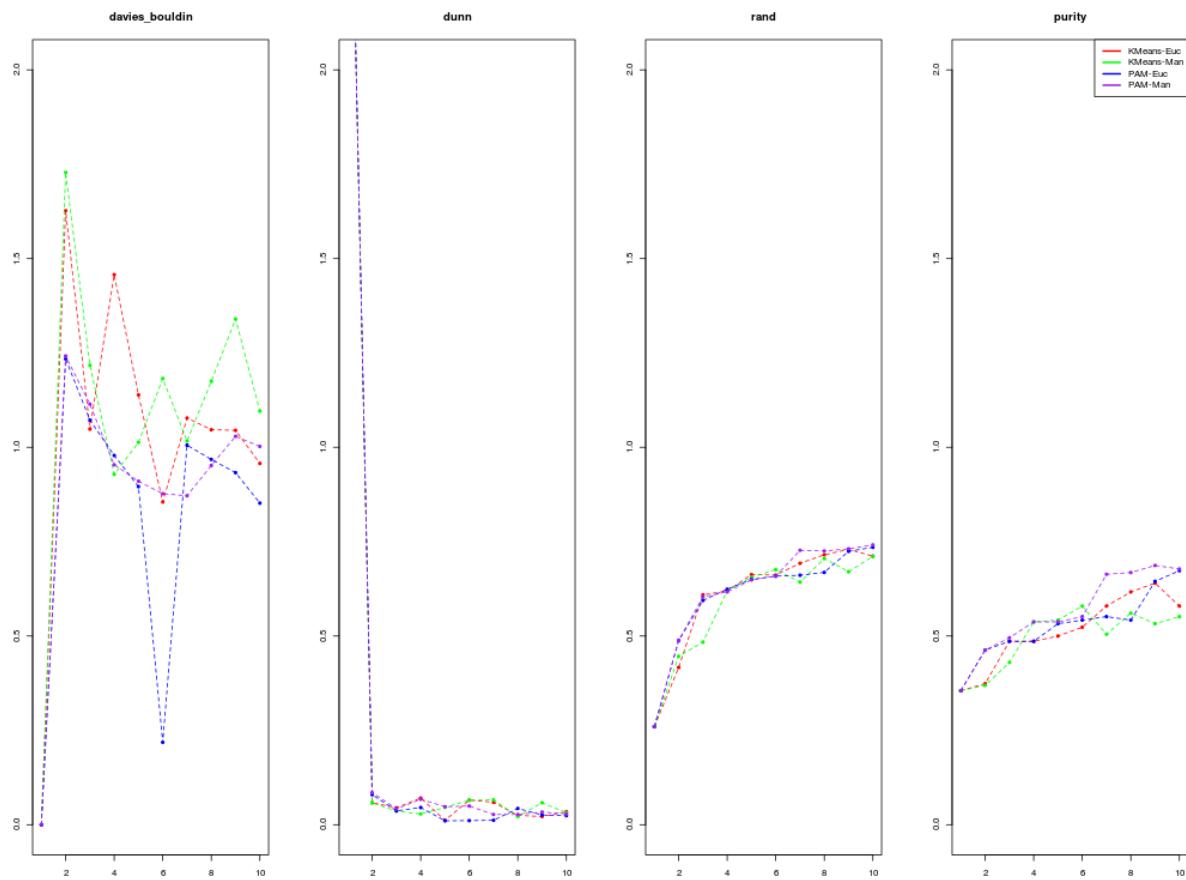


Rysunek 8: Wynik klasteryzacji dla algorytmu PAM (Manhattan) dla zbioru "Diabetes".



Rysunek 9: Klastry dla algorytmu PAM (Manhattan) dla zbioru "Diabetes".

2.2 Zbiór "Glass"

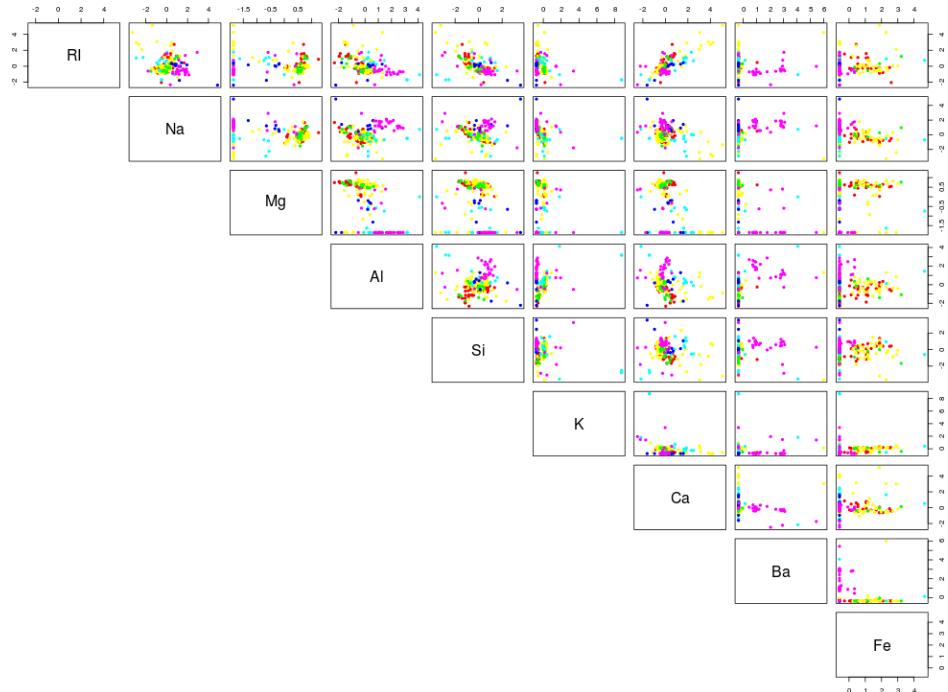


Rysunek 10: Wykresy wartości metryk dla zbioru "Glass".

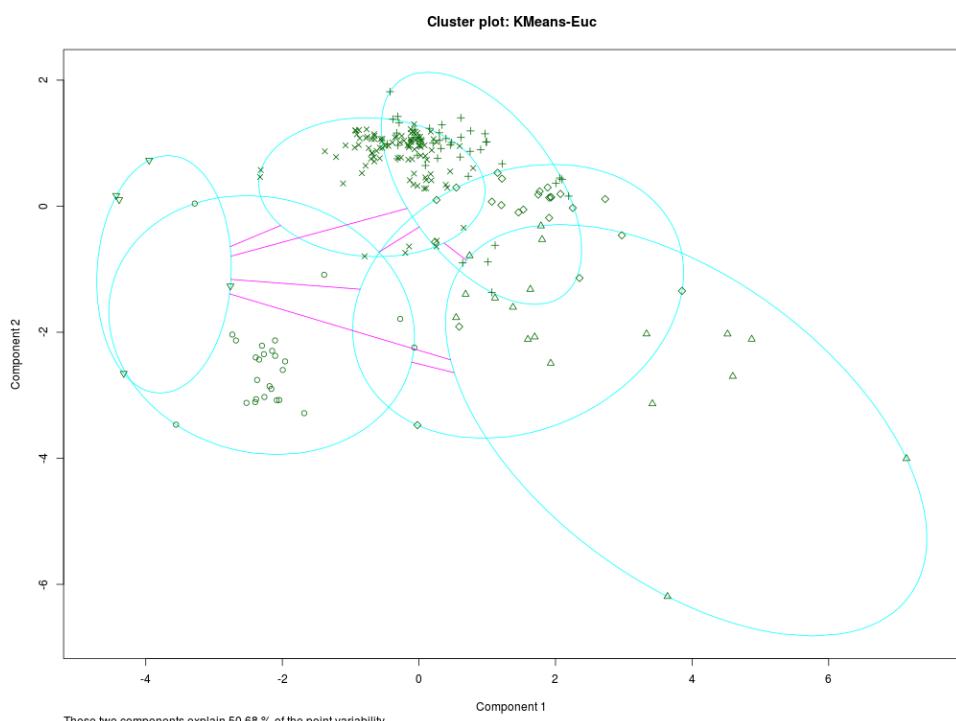
Algorytm	Liczba klastrów	davies_bouldin	dunn	rand	purity
KMeans-Euc	1	0.00	INF	0.26	0.35
	2	1.63	0.058	0.42	0.37
	3	1.05	0.045	0.61	0.49
	4	1.46	0.071	0.62	0.49
	5	1.14	0.012	0.66	0.50
	6	0.85	0.065	0.66	0.52
	7	1.08	0.06	0.69	0.58
	8	1.05	0.025	0.71	0.62
	9	1.04	0.022	0.73	0.64
	10	0.96	0.035	0.71	0.58
KMeans-Man	1	0.00	INF	0.26	0.35
	2	1.73	0.058	0.45	0.37
	3	1.22	0.036	0.48	0.43
	4	0.93	0.029	0.62	0.54
	5	1.01	0.046	0.66	0.54
	6	1.18	0.065	0.68	0.58
	7	1.02	0.066	0.64	0.51
	8	1.18	0.022	0.70	0.56
	9	1.34	0.058	0.67	0.53
	10	1.10	0.032	0.71	0.55
PAM-Euc	1	0.00	INF	0.26	0.35
	2	1.23	0.08	0.49	0.46
	3	1.07	0.037	0.59	0.49
	4	0.98	0.046	0.62	0.49
	5	0.90	0.01	0.65	0.53
	6	0.22	0.011	0.66	0.54
	7	1.01	0.012	0.66	0.55
	8	0.97	0.043	0.67	0.54
	9	0.93	0.026	0.72	0.65
	10	0.85	0.024	0.73	0.67
PAM-Man	1	0.00	INF	0.26	0.35
	2	1.24	0.086	0.49	0.46
	3	1.11	0.042	0.60	0.49
	4	0.95	0.067	0.62	0.54
	5	0.91	0.048	0.65	0.54
	6	0.88	0.05	0.66	0.55
	7	0.87	0.027	0.73	0.66
	8	0.95	0.027	0.72	0.67
	9	1.03	0.034	0.73	0.69
	10	1.00	0.029	0.74	0.68

Tabela 2: Wartości metryk dla zbioru "Glass".

2.2.1 Algorytm K-Means (Euclidean)

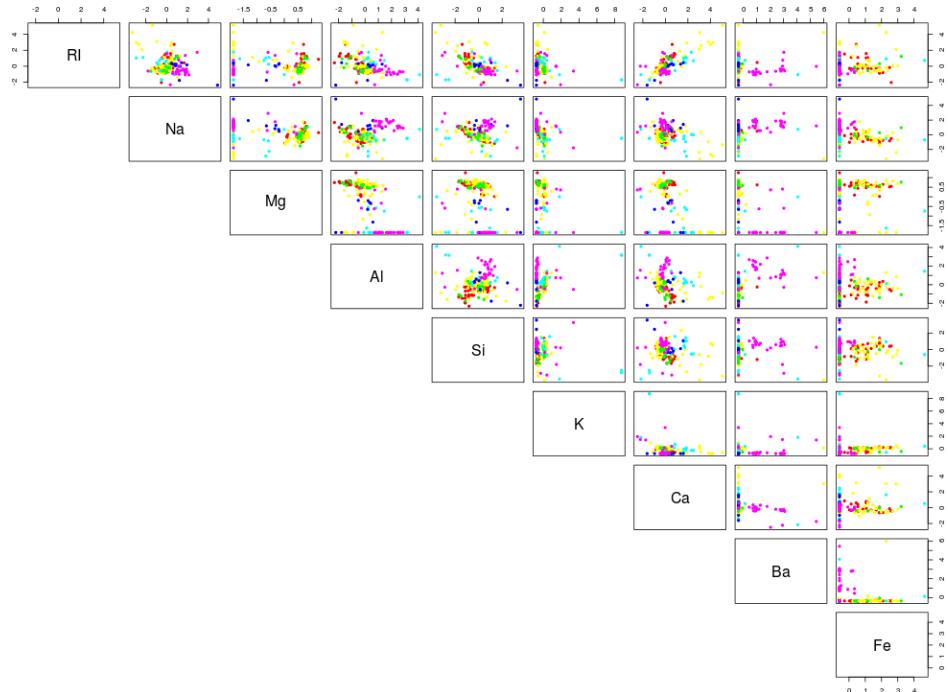


Rysunek 11: Wynik klasteryzacji dla algorytmu KMeans (Euclidean) dla zbioru "Glass".

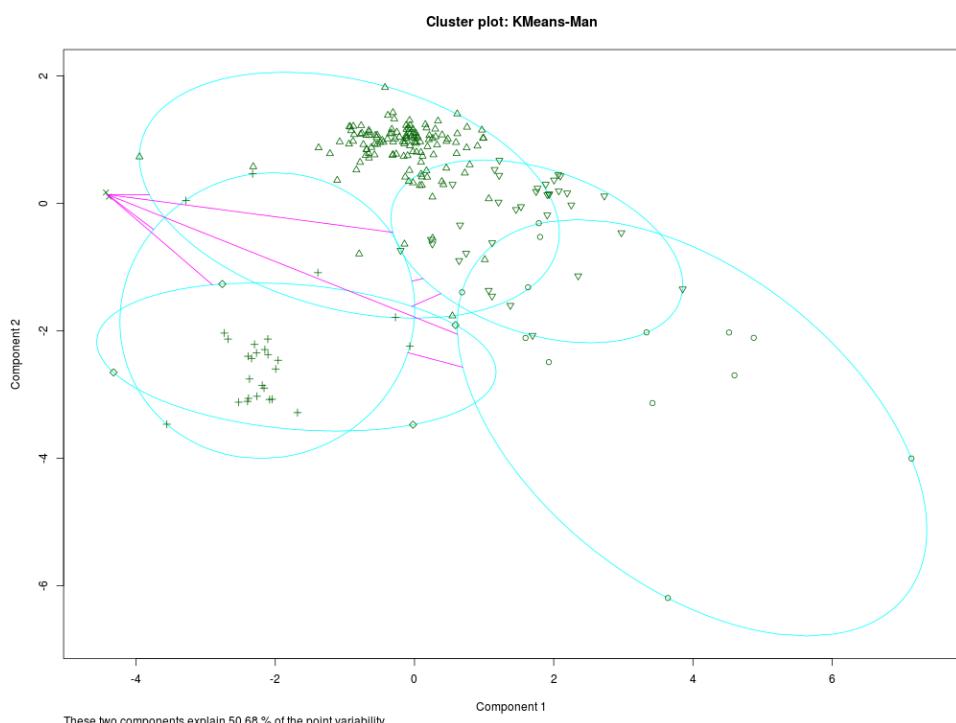


Rysunek 12: Klastry dla algorytmu KMeans (Euclidean) dla zbioru "Glass".

2.2.2 Algorytm K-Means (Manhattan)

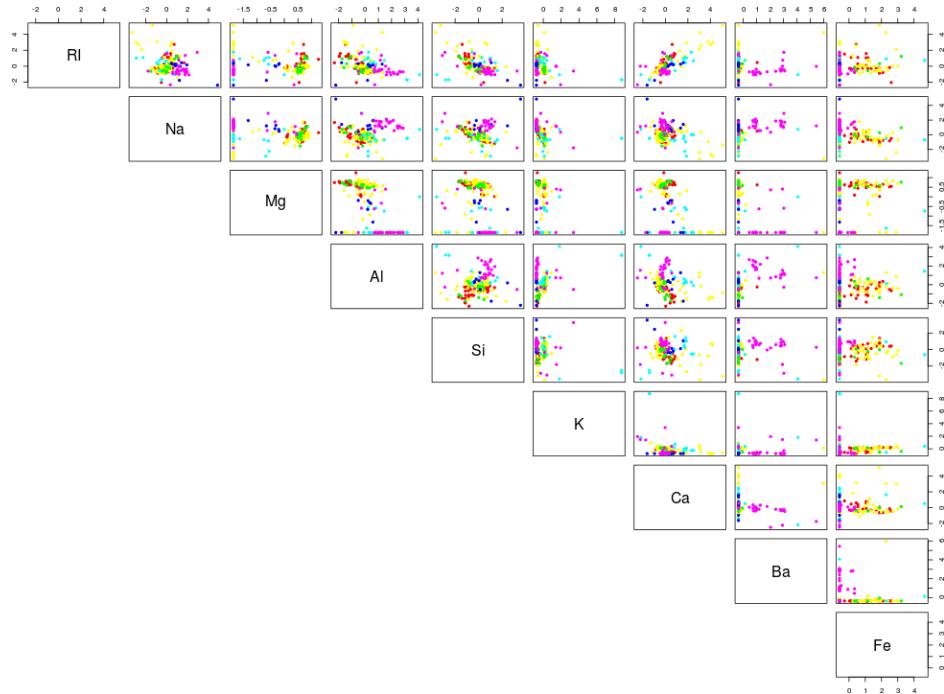


Rysunek 13: Wynik klasteryzacji dla algorytmu KMeans (Manhattan) dla zbioru "Glass".

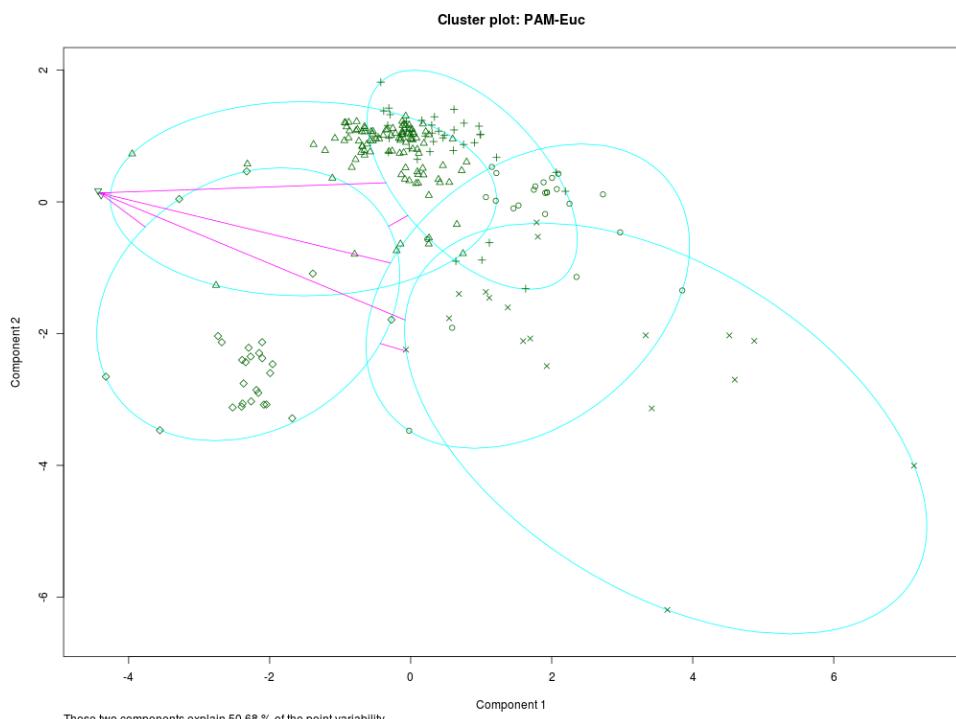


Rysunek 14: Klastry dla algorytmu KMeans (Manhattan) dla zbioru "Glass".

2.2.3 Algorytm PAM (Euclidean)

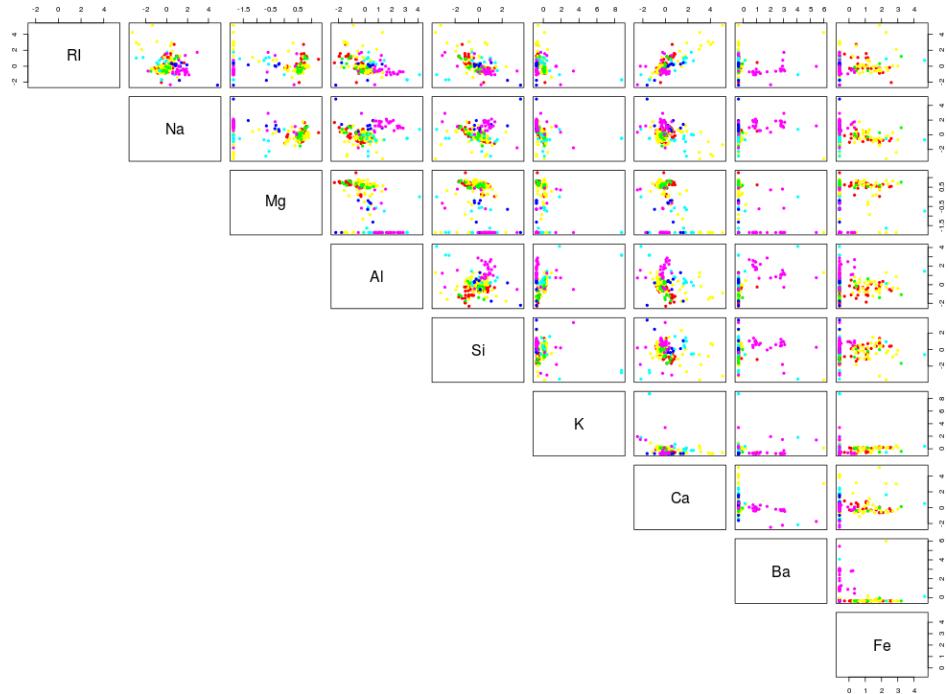


Rysunek 15: Wynik klasteryzacji dla algorytmu PAM (Euclidean) dla zbioru "Glass".

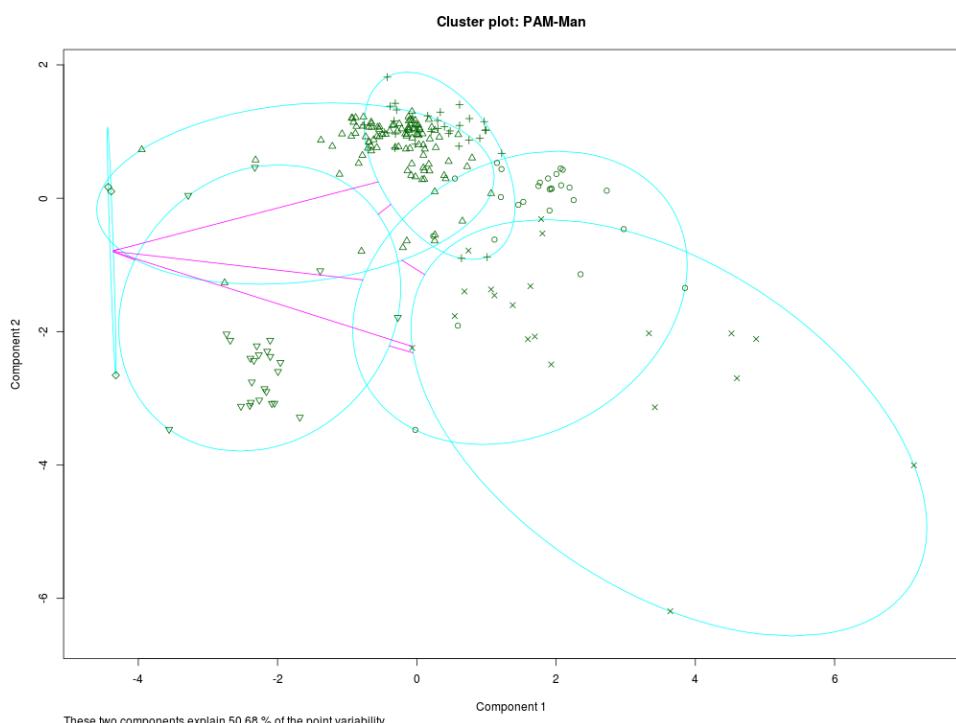


Rysunek 16: Klastry dla algorytmu PAM (Euclidean) dla zbioru "Glass".

2.2.4 Algorytm PAM (Manhattan)

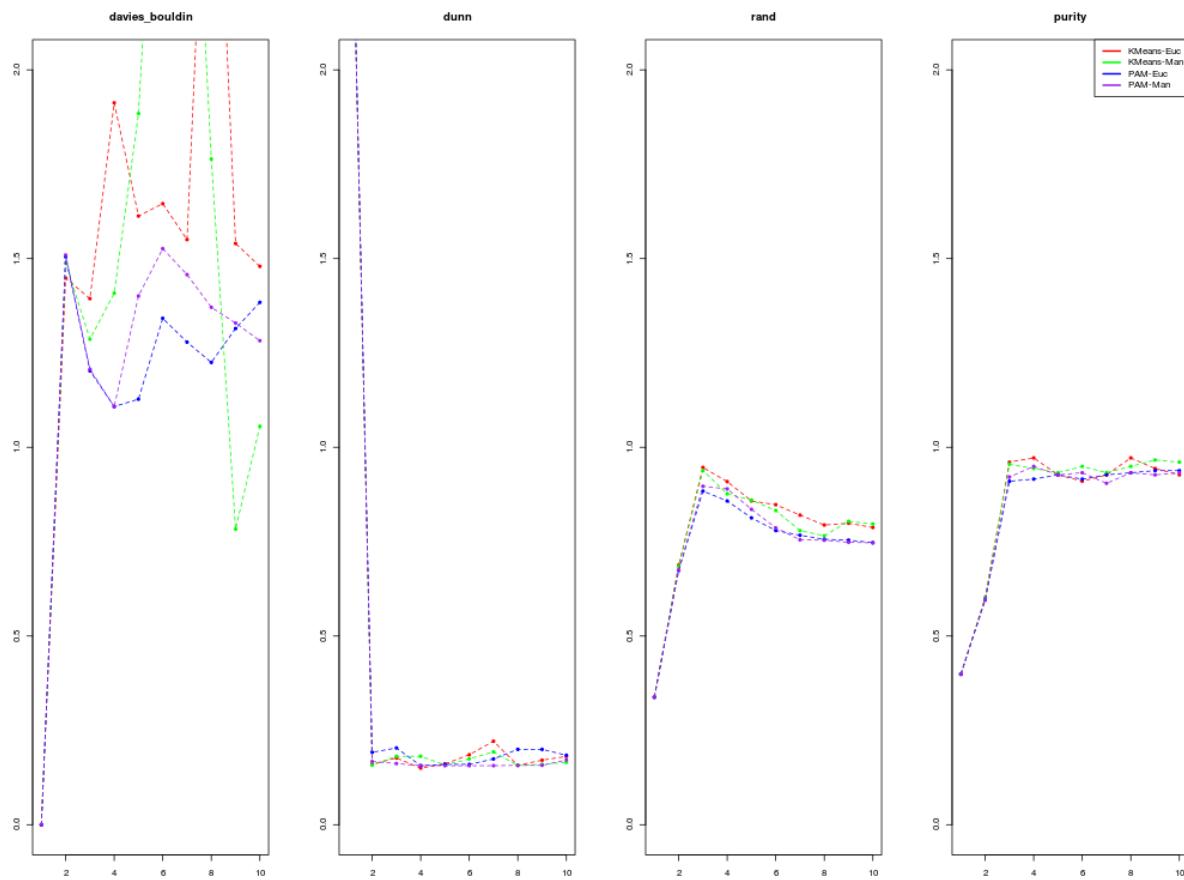


Rysunek 17: Wynik klasteryzacji dla algorytmu PAM (Manhattan) dla zbioru "Glass".



Rysunek 18: Klastry dla algorytmu PAM (Manhattan) dla zbioru "Glass".

2.3 Zbiór "Wine"

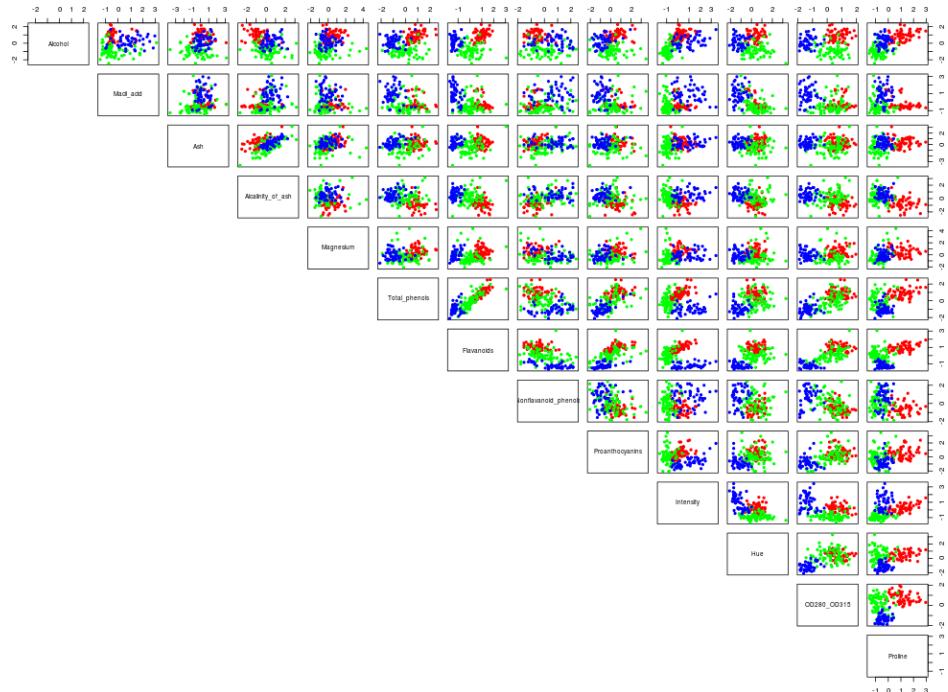


Rysunek 19: Wykresy wartości metryk dla zbioru "Wine".

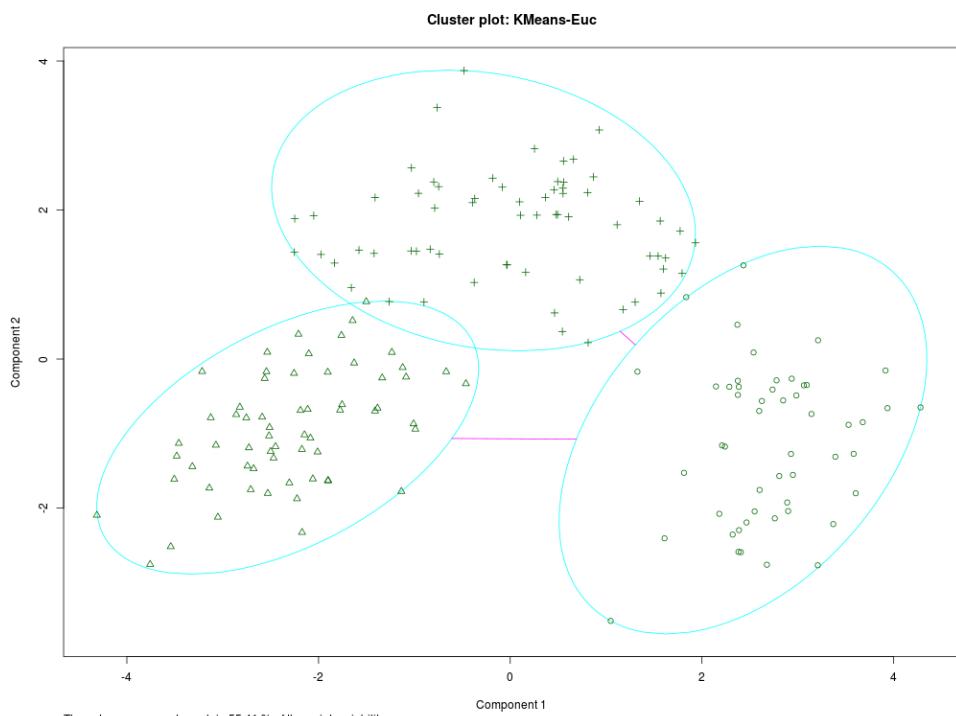
Algorytm	Liczba klastrów	davies_bouldin	dunn	rand	purity
KMeans-Euc	1	0.00	INF	0.34	0.40
	2	1.45	0.16	0.69	0.60
	3	1.39	0.177	0.95	0.96
	4	1.91	0.15	0.91	0.97
	5	1.61	0.161	0.86	0.93
	6	1.65	0.185	0.85	0.91
	7	1.55	0.221	0.82	0.93
	8	2.08	0.157	0.79	0.97
	9	1.54	0.171	0.80	0.94
	10	1.48	0.181	0.79	0.93
KMeans-Man	1	0.00	INF	0.34	0.40
	2	1.49	0.158	0.69	0.60
	3	1.29	0.181	0.94	0.95
	4	1.41	0.181	0.88	0.94
	5	1.88	0.158	0.86	0.93
	6	2.06	0.174	0.83	0.95
	7	2.07	0.193	0.78	0.93
	8	1.76	0.157	0.77	0.95
	9	0.78	0.16	0.80	0.97
	10	1.06	0.165	0.80	0.96
PAM-Euc	1	0.00	INF	0.34	0.40
	2	1.50	0.192	0.67	0.60
	3	1.20	0.203	0.88	0.91
	4	1.11	0.156	0.86	0.92
	5	1.13	0.16	0.81	0.93
	6	1.34	0.16	0.78	0.92
	7	1.28	0.174	0.77	0.93
	8	1.23	0.2	0.76	0.93
	9	1.31	0.2	0.75	0.94
	10	1.38	0.184	0.75	0.94
PAM-Man	1	0.00	INF	0.34	0.40
	2	1.51	0.167	0.67	0.60
	3	1.21	0.162	0.90	0.92
	4	1.11	0.156	0.89	0.95
	5	1.40	0.156	0.84	0.93
	6	1.53	0.156	0.79	0.93
	7	1.46	0.156	0.76	0.90
	8	1.37	0.157	0.75	0.93
	9	1.33	0.157	0.75	0.93
	10	1.28	0.172	0.75	0.93

Tabela 3: Wartości metryk dla zbioru "Wine".

2.3.1 Algorytm K-Means (Euclidean)

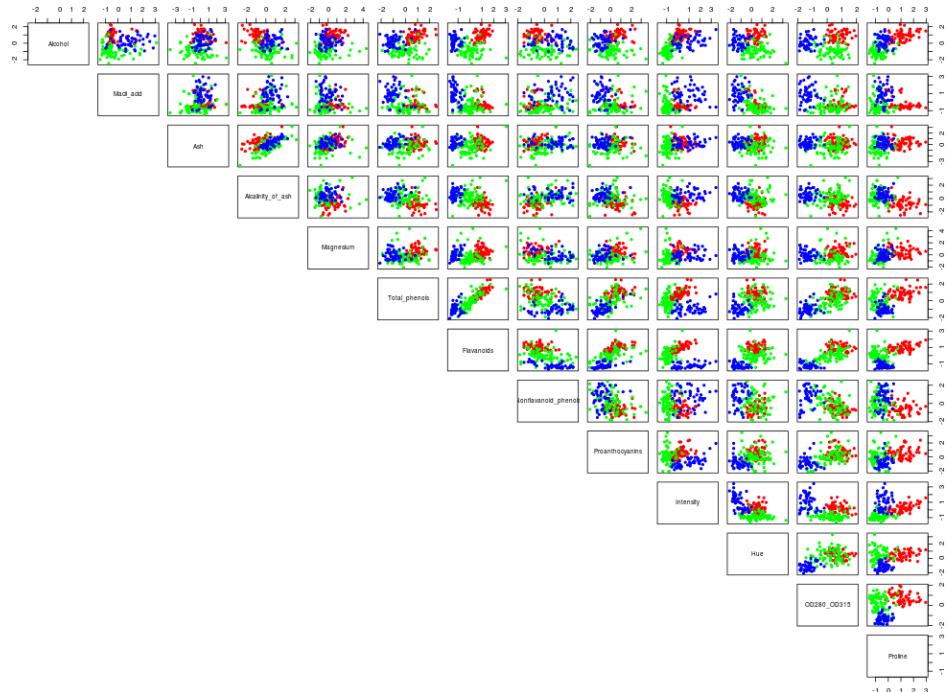


Rysunek 20: Wynik klasteryzacji dla algorytmu KMeans (Euclidean) dla zbioru "Wine".

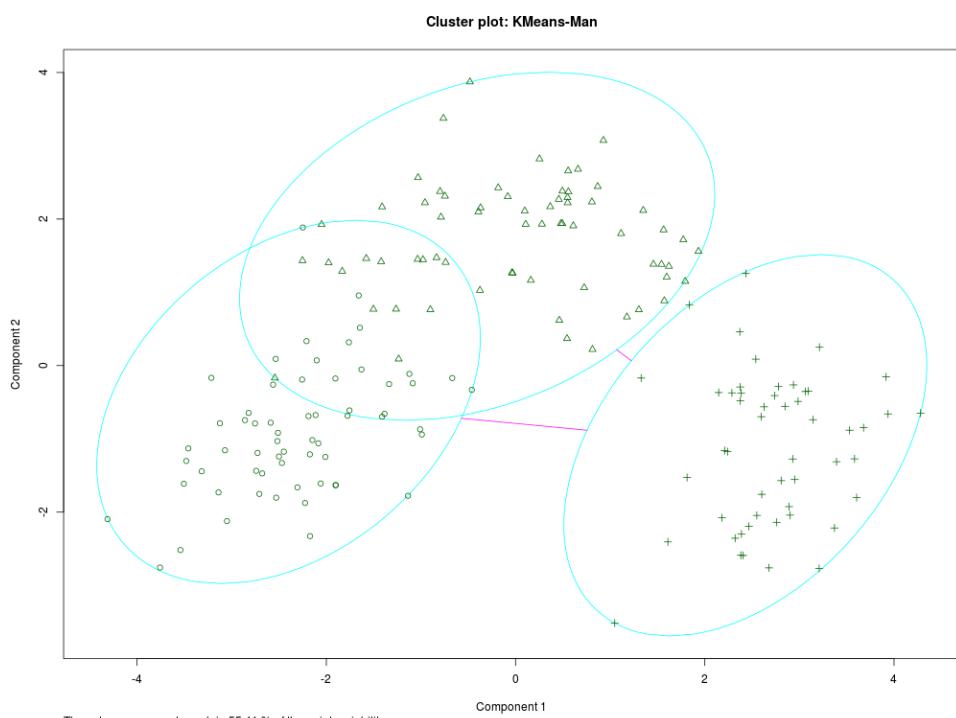


Rysunek 21: Klastry dla algorytmu KMeans (Euclidean) dla zbioru "Wine".

2.3.2 Algorytm K-Means (Manhattan)

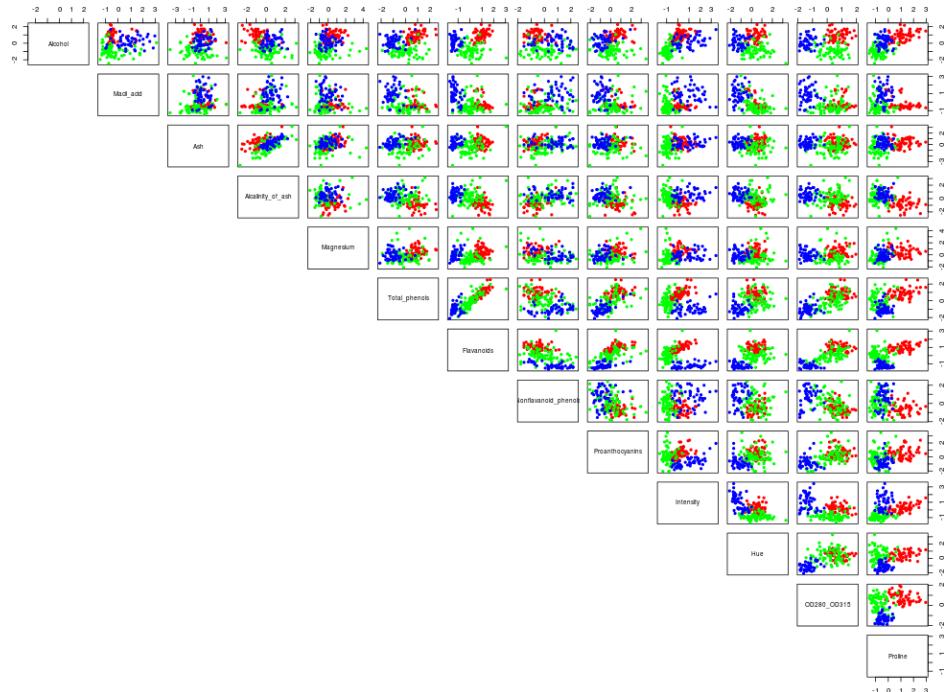


Rysunek 22: Wynik klasteryzacji dla algorytmu KMeans (Manhattan) dla zbioru "Wine".

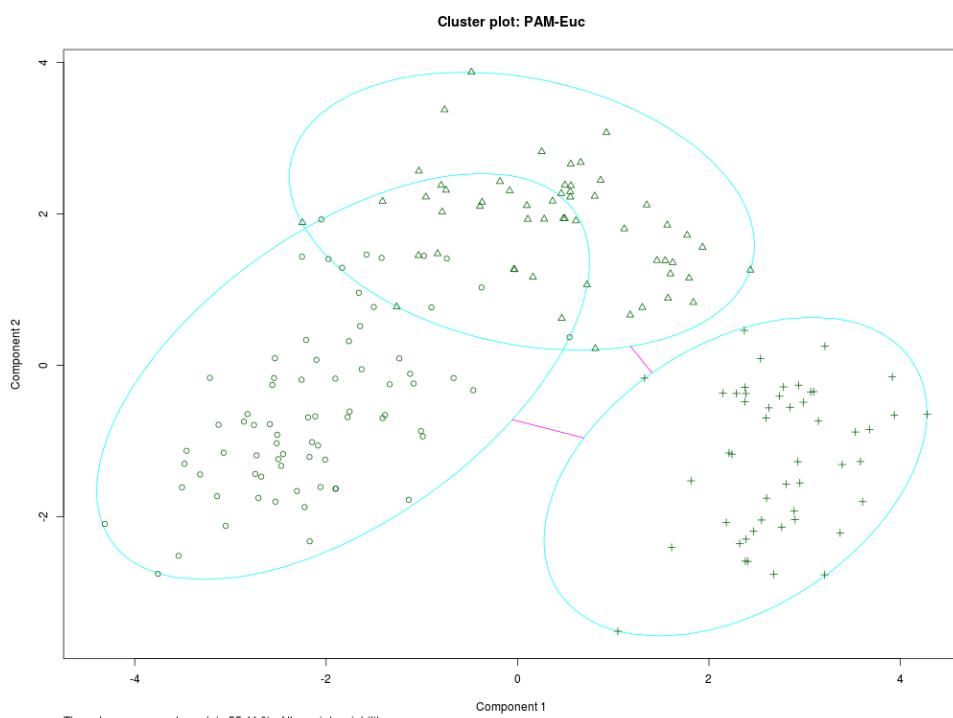


Rysunek 23: Klastry dla algorytmu KMeans (Manhattan) dla zbioru "Wine".

2.3.3 Algorytm PAM (Euclidean)

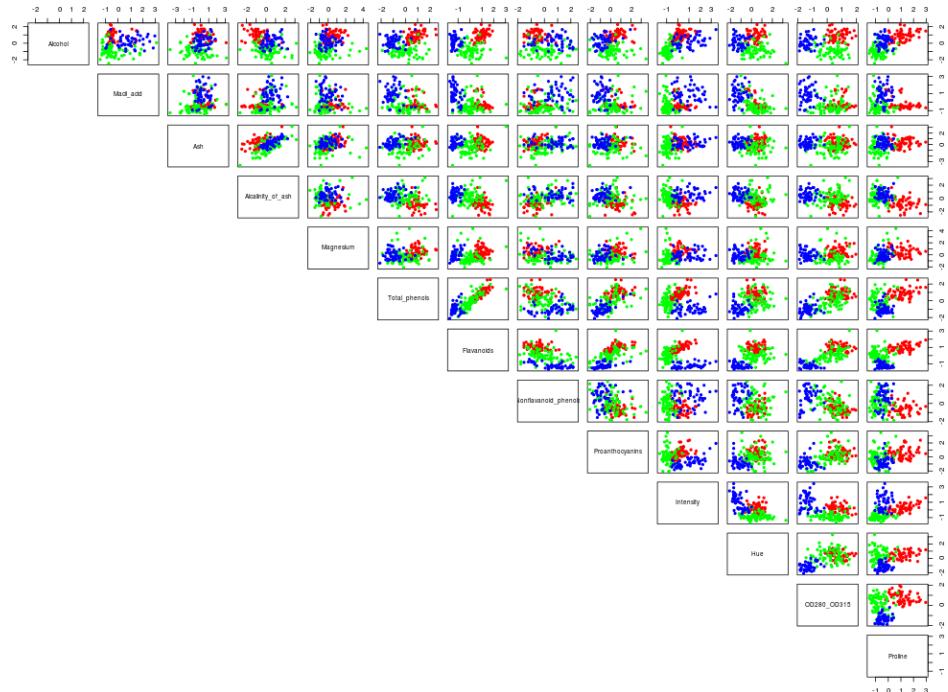


Rysunek 24: Wynik klasteryzacji dla algorytmu PAM (Euclidean) dla zbioru "Wine".

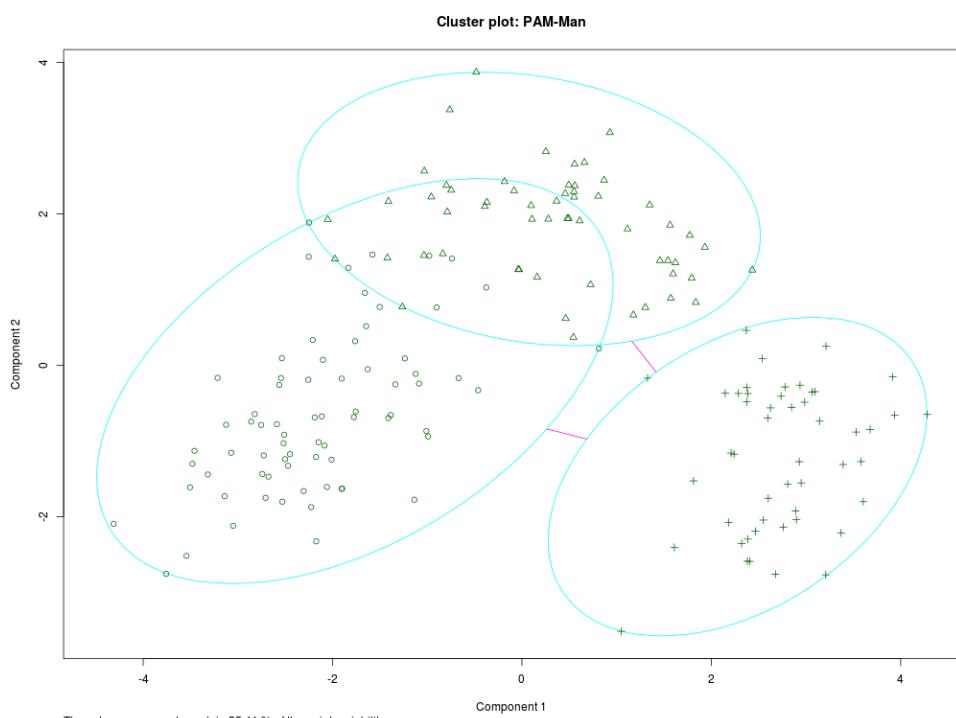


Rysunek 25: Klastry dla algorytmu PAM (Euclidean) dla zbioru "Wine".

2.3.4 Algorytm PAM (Manhattan)

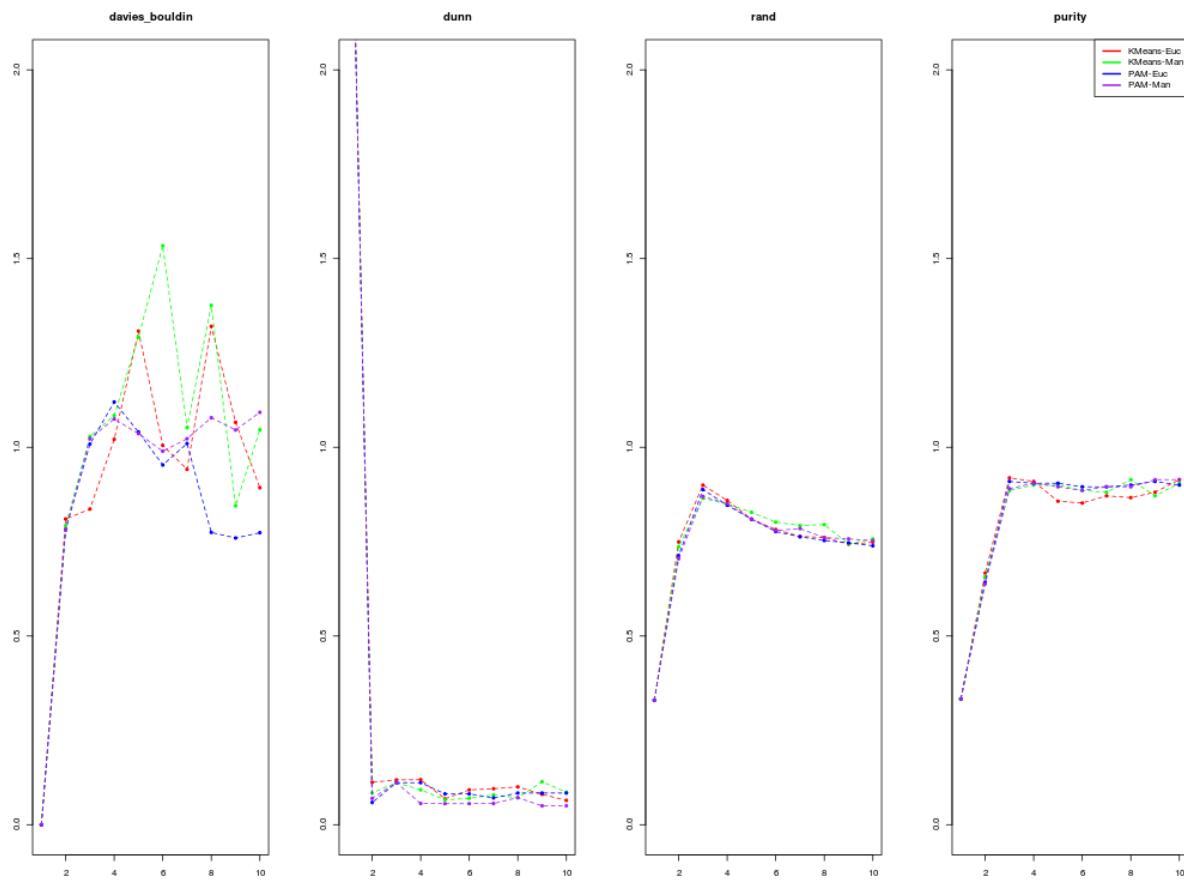


Rysunek 26: Wynik klasteryzacji dla algorytmu PAM (Manhattan) dla zbioru "Wine".



Rysunek 27: Klastry dla algorytmu PAM (Manhattan) dla zbioru "Wine".

2.4 Zbiór "Seeds"

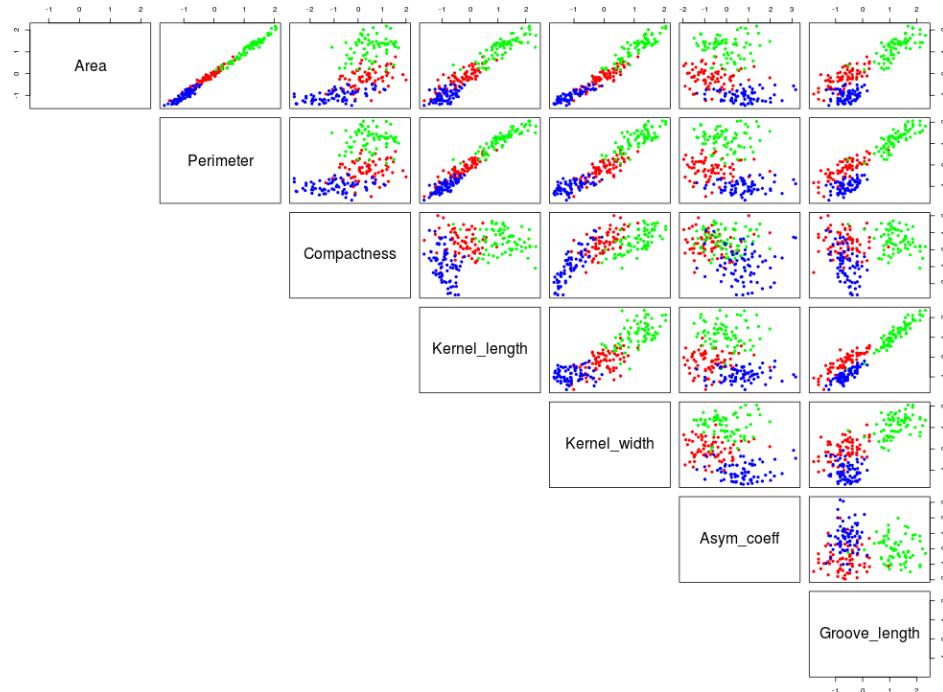


Rysunek 28: Wykresy wartości metryk dla zbioru "Seeds".

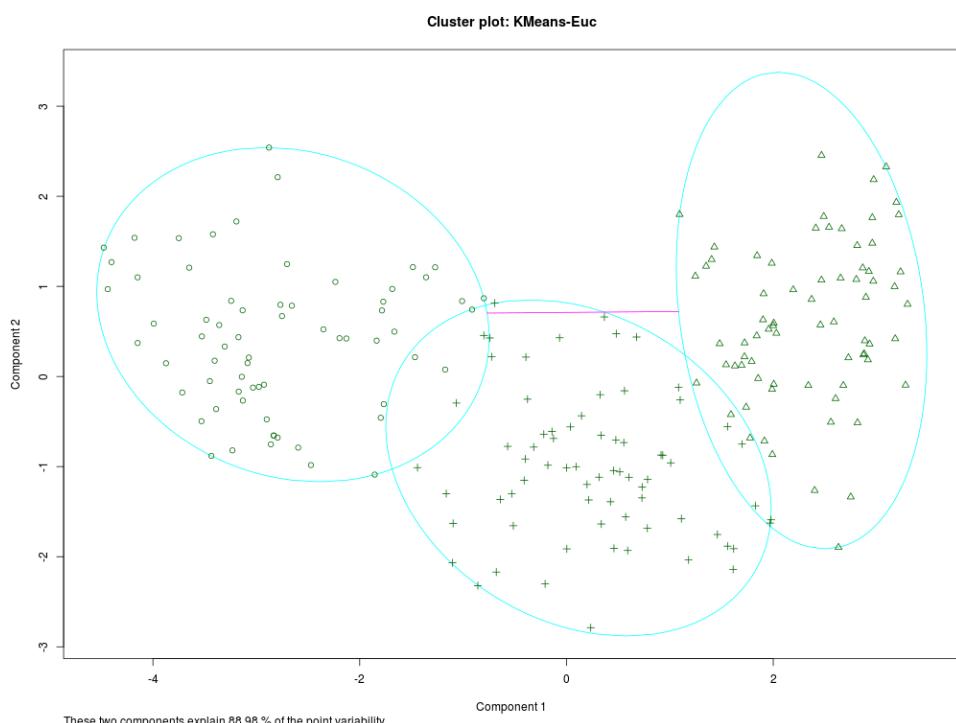
Algorytm	Liczba klastrów	davies_bouldin	dunn	rand	purity
KMeans-Euc	1	0.00	INF	0.33	0.33
	2	0.81	0.113	0.75	0.67
	3	0.84	0.119	0.90	0.92
	4	1.02	0.119	0.86	0.91
	5	1.31	0.069	0.81	0.86
	6	1.00	0.092	0.78	0.85
	7	0.94	0.095	0.76	0.87
	8	1.32	0.1	0.76	0.87
	9	1.07	0.081	0.74	0.88
	10	0.89	0.064	0.75	0.91
KMeans-Man	1	0.00	INF	0.33	0.33
	2	0.79	0.084	0.74	0.66
	3	1.03	0.111	0.86	0.89
	4	1.08	0.092	0.85	0.90
	5	1.29	0.066	0.83	0.90
	6	1.53	0.07	0.80	0.89
	7	1.05	0.079	0.79	0.88
	8	1.38	0.071	0.80	0.91
	9	0.84	0.114	0.74	0.87
	10	1.05	0.086	0.76	0.91
PAM-Euc	1	0.00	INF	0.33	0.33
	2	0.78	0.059	0.71	0.64
	3	1.01	0.111	0.89	0.91
	4	1.12	0.111	0.85	0.91
	5	1.04	0.082	0.81	0.91
	6	0.95	0.082	0.78	0.90
	7	1.01	0.071	0.76	0.90
	8	0.77	0.084	0.75	0.90
	9	0.76	0.084	0.75	0.91
	10	0.77	0.084	0.74	0.90
PAM-Man	1	0.00	INF	0.33	0.33
	2	0.78	0.07	0.70	0.64
	3	1.02	0.111	0.87	0.89
	4	1.07	0.056	0.85	0.91
	5	1.04	0.056	0.81	0.90
	6	0.99	0.056	0.78	0.89
	7	1.02	0.056	0.79	0.90
	8	1.08	0.072	0.76	0.90
	9	1.05	0.05	0.76	0.91
	10	1.09	0.05	0.75	0.91

Tabela 4: Wartości metryk dla zbioru "Seeds".

2.4.1 Algorytm K-Means (Euclidean)

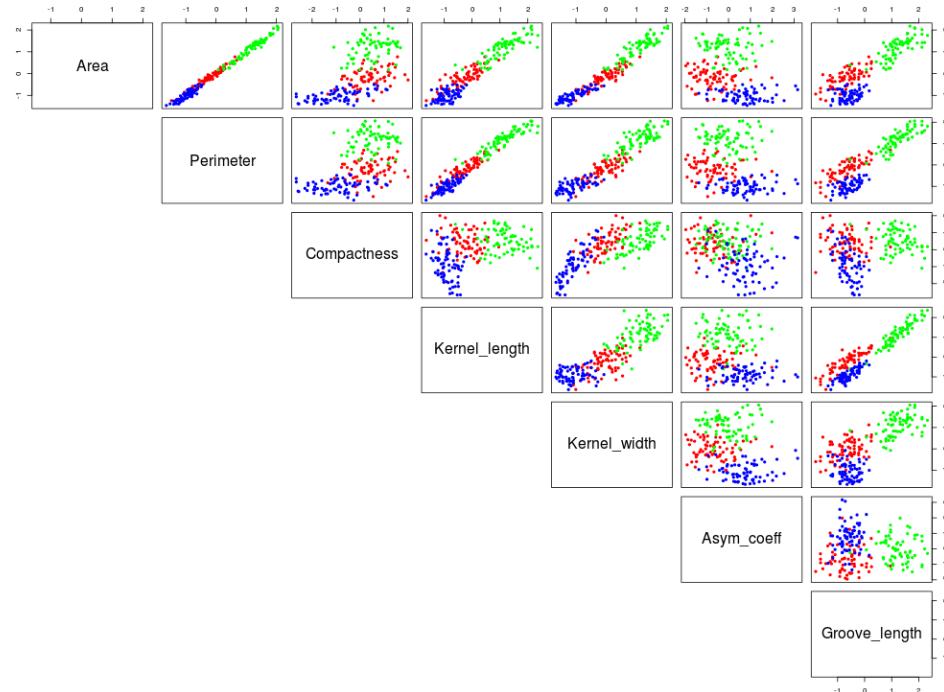


Rysunek 29: Wynik klasteryzacji dla algorytmu KMeans (Euclidean) dla zbioru "Seeds".

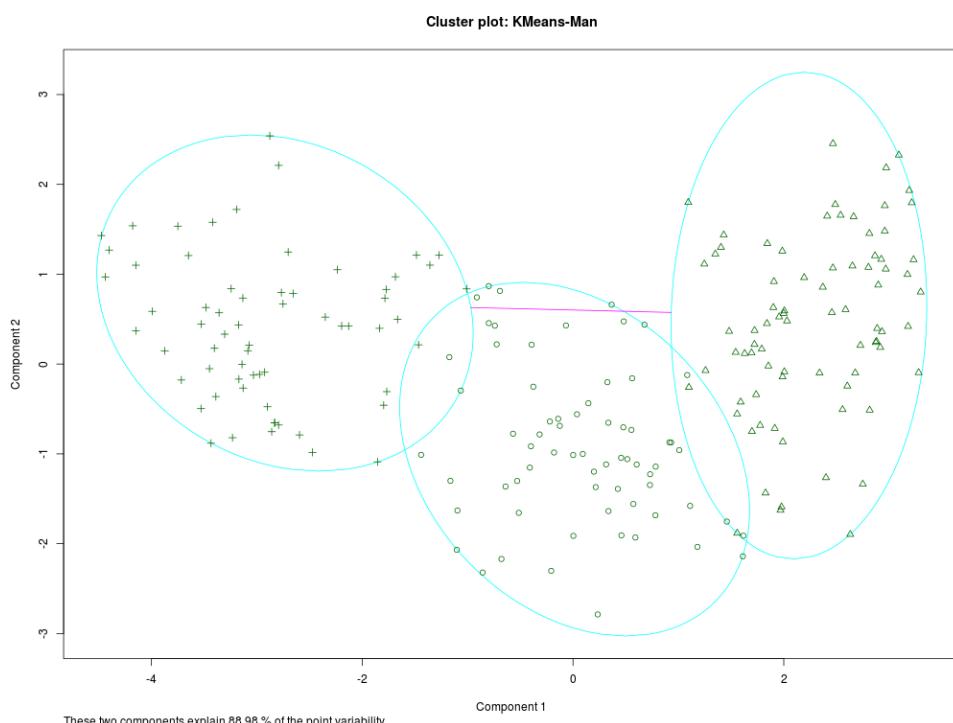


Rysunek 30: Klastry dla algorytmu KMeans (Euclidean) dla zbioru "Seeds".

2.4.2 Algorytm K-Means (Manhattan)

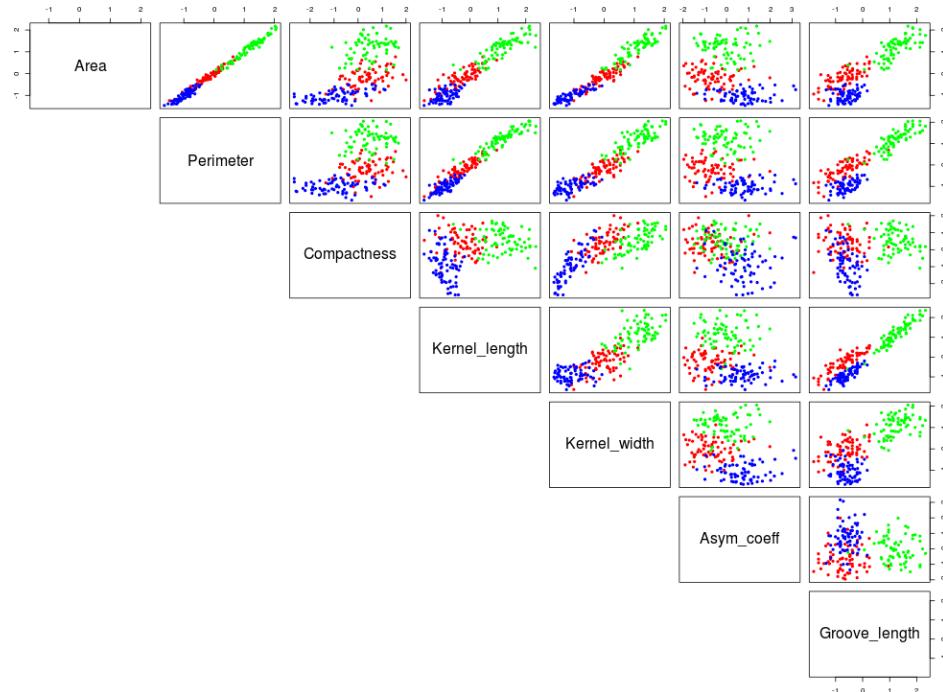


Rysunek 31: Wynik klasteryzacji dla algorytmu KMeans (Manhattan) dla zbioru "Seeds".

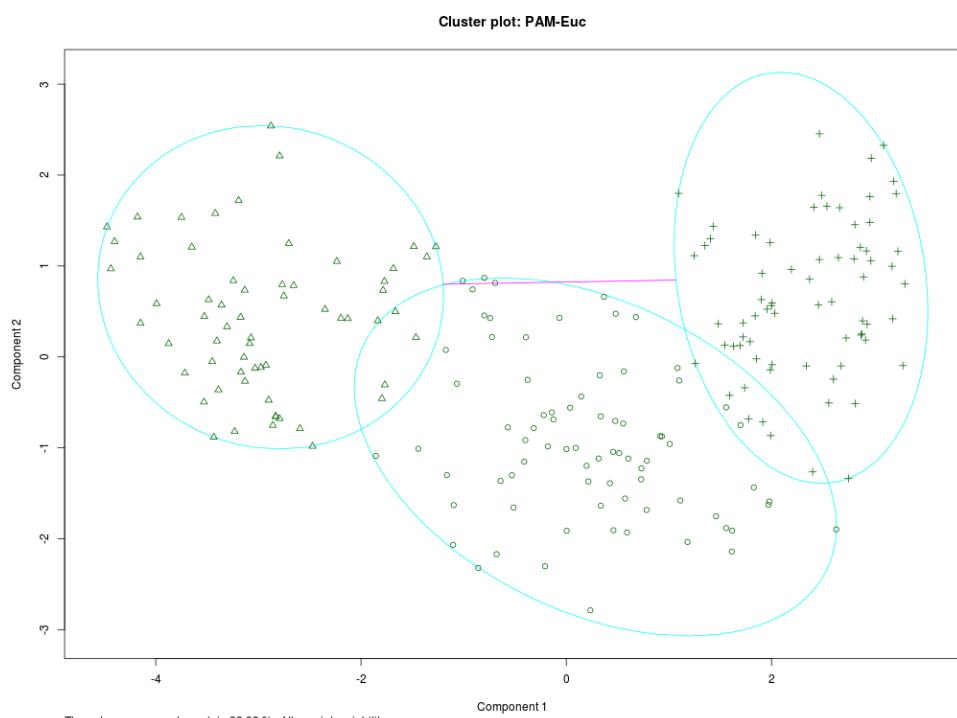


Rysunek 32: Klastry dla algorytmu KMeans (Manhattan) dla zbioru "Seeds".

2.4.3 Algorytm PAM (Euclidean)

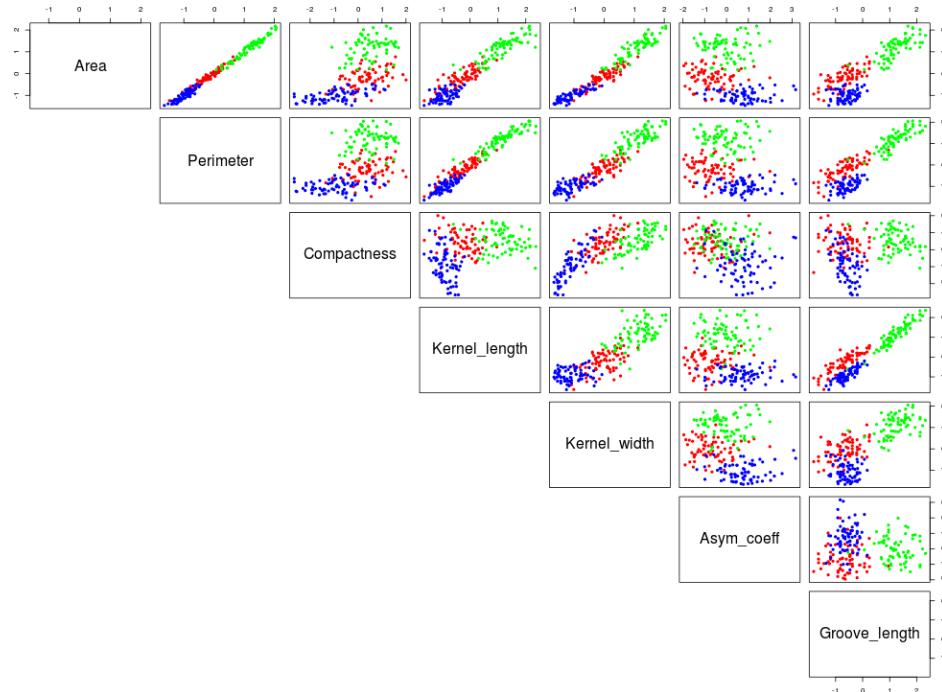


Rysunek 33: Wynik klasteryzacji dla algorytmu PAM (Euclidean) dla zbioru "Seeds".

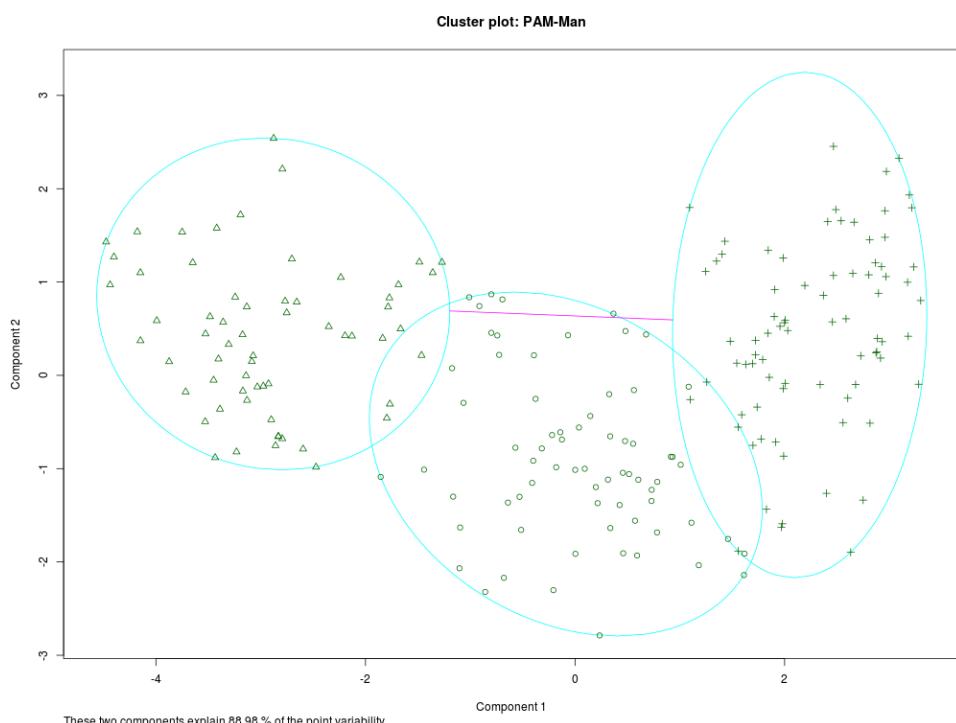


Rysunek 34: Klastry dla algorytmu PAM (Euclidean) dla zbioru "Seeds".

2.4.4 Algorytm PAM (Manhattan)



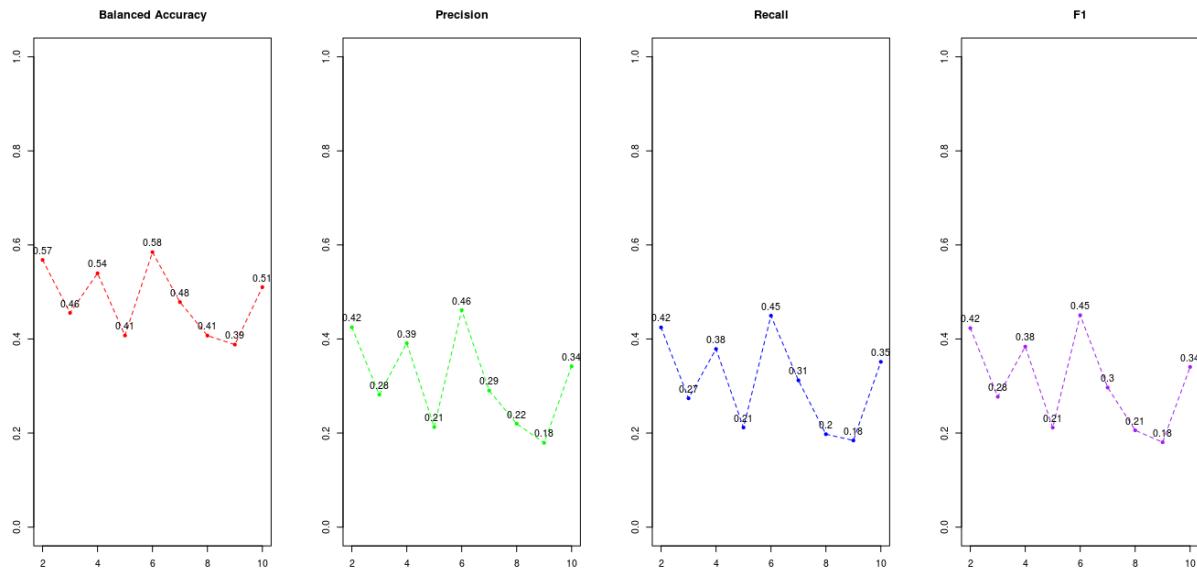
Rysunek 35: Wynik klasteryzacji dla algorytmu PAM (Manhattan) dla zbioru "Seeds".



Rysunek 36: Klastry dla algorytmu PAM (Manhattan) dla zbioru "Seeds".

2.5 Kroswalidacja Seeds

Przeprowadzono kroswalidację dla zbioru *Seeds* z użyciem algorytmu Kmeans z parametrami: liczba grup = 3 (liczba klas), metryka euklidesowa. Zbadano wartości miar *Balanced Accuracy*, *Precision*, *Recall* oraz *F1* dla kroswalidacji kolejno 2, 3, ..., 10-fold. Wyniki pomiarów zostały przedstawione poniżej:



Rysunek 37: Wyniki kroswalidacji dla zbioru *Seeds*.