Szymon Gruda	239661	239661@edu.p.lodz.pl
Jan Karwowski	239671	239671@edu.p.lodz.pl
Michał Kidawa	239673	239673@edu.p.lodz.pl
Kamil Kowalewski	239676	239676@edu.p.lodz.pl

Zadanie 4.: Problem Set 4

Spis treści

1. Wprowadzenie

1.1. Metoda k-średnich

Na początku zbadana wartość współczynnika zarysu (ang. silhouette coefficient) dla każdego zbioru w zależności od liczby klastrów (przyjęto, że będzie to liczba naturalna z przedziału [2; 30]), w tym przypadku ustawiono, zgodnie z konwencją, liczbę 300 jako maksymalną liczbę iteracji. Następnie na podstawie uzyskanych wyników wybrano pięcioelementowy zbiór wartości parametru liczby klastrów, dla którego zbadano zachowanie współczynnika zarysu w zależności od maksymalnej liczby iteracji dla każdego zbioru.

1.2. Metoda aglomeracyjna

Jeżeli chodzi o algorytm aglomeracyjny, to jest to algorytm hierarchiczny, który na początku zakłada, że każda próbka jest osobnym klastrem. W kolejnych iteracjach łączy najbliższe siebie klastry w jeden. W ten sposób z każdą iteracją jest jeden klaster mniej. Algorytm kończy się, kiedy powstanie zadana liczba klastrów, w skrajnym przypadku jeden. Podobieństwo klastrów mierzy się z wykorzystaniem odpowiedniej metryki (np. euklidesowej bądź miejskiej) i odpowiedniej metody łączenia. Te dwa parametry mają zasadniczy wpływ na działanie algorytmu. Trzecim badanym parametrem jest liczba klastrów, na której algorytm się zatrzymuje. Oczywiście aby osiągnąć zadaną liczbę klastrów K dla N próbek, algorytm aglomeracyjny musi przejść przez wszystkie całkowite liczby klastrów od N do K.

1.3. Metoda EM

Algorytm Oczekiwania-Maksymalizacji (ang. *Expectation-Maximization*, w ramach sprawozdania *EM*) to iteracyjna metoda, która wylicza rozkłady prawdopodobieństwa przynależności obserwacji do danego centrum. Metoda składa się z dwóch kroków wykonywanych na przemian w iteracjach, aż do osiągnięcia warunku końcowego:

— estymacja (expectation) - dla wyestymowanego układu parametrów rozkładu przypadków dokonywane jest przypisanie przykładom prawdopodobieństwa przynależenia do skupień
— maksymalizacja (maximization) - wyznaczenie wartości parametrów skupień, przy których wiarygodność rozkładu osiąga maksymalną wartość

Parametry, które ulegały zmianie podczas eksperymentów to liczba maksymalnych iteracji, typ macierzy kowariancji oraz liczba klastrów.

1.4. Metoda DBSCAN

Na początku warto zaznaczyć, że dla metody DBSCAN zakres wartości eps dla zbioru[1] oraz [3] był *0.1 do 10* z częstością co *0.1* natomiast ze względu na charakterystykę zbioru [2] musiał zostać użyty inny przedział aby uzyskać najlepsze wyniki a mianowicie *15 do 30* z częstością co *0.1*. Z tego powodu, że wyników było bardzo dużo w tabelach dla zbioru [2] zostały one pominięte.

1.5. Opis zbiorów danych

1.5.1. Zbiór Iris Species

Jeden z najpopularniejszych zbiorów wykorzystywanych w Machine Learningu. Zbiór zawiera informacje o 3 gatunkach Irysów, po 50 próbek na gatunek.Jeden gatunek jest liniowo separowalny od pozostałych dwóch, które nie są liniowo separowalne od siebie nawzajem. Zbiór składa się z następujących kolumn:
— *ID* - identyfikator;
— *SepalLengthCm* - długość działki kielicha w cm;
— *SepalWidthCm* - szerokość działki kielicha w cm;
— *PetalLengthCm* - długość płatka w cm;
— *PetalWidthCm* - szerokość płatka w cm;
— *Species* - nazwa gatunku;

1.5.2. Zbiór Mall Customer Segmentation Data

Zbiór zawiera informacje o klientach centrum handlowego, zebrane przy okazji zakładania przez klientów kont członkowskich. Zbiór zawiera 200 rekordów i składa się z następujących kolumn:
— *CustomerID* - identyfikator;
— *Gender* - płeć;
— *Age* - wiek;
— *Annual Income* - roczny dochód w dolarach amerykańskich;
— *Spending Score* - ocena klienta przyznana na podstawie jego zachowań oraz wydatków;

1.5.3. Zbiór Moons

Ostatnim z wybranych zbiorów był dostępny w ramach pakietu sklearn zbiór *moons*. Za pomocą metody generowana jest podana liczba próbek w postaci tupli o dwóch wartościach. Dataset jest opisywany jako przeznaczony do testowania algorytmów klasteryzacji i klasyfikacji. Podczas eksperymentów przeprowadzonych w ramach realizacji zadania dla każdego algorytmu generujemy 700 próbek.

2. Wyniki

2.1. Algorytm k-średnich

Clusters	Silhouette
2	0.6808
3	0.5526
4	0.4978
5	0.4885
6	0.3677
7	0.356
8	0.3552
9	0.3449
10	0.3163
11	0.3177
12	0.2961
13	0.2936
14	0.3134
15	0.2914
16	0.3004
17	0.2891
18	0.2648
19	0.2703
20	0.2755
21	0.2824
22	0.2803
23	0.2732
24	0.2821
25	0.2712
26	0.2571
27	0.281
28	0.2567
29	0.271
30	0.2726

Tabela 1. Wartość Silhouette w zależności od liczby klastrów dla zbioru Iris Species

Clusters	Silhouette
2	0.2931
3	0.3838
4	0.4053
5	0.4441
6	0.4514
7	0.4395
8	0.4305
9	0.3915
10	0.3843
11	0.3706
12	0.3504
13	0.3523
14	0.3389
15	0.3507
16	0.3468
17	0.3391
18	0.3163
19	0.3189
20	0.3356
21	0.336
22	0.3362
23	0.3371
24	0.3366
25	0.3266
26	0.3309
27	0.3338
28	0.3262
29	0.3461
30	0.3384

Tabela 2. Wartość Silhouette w zależności od liczby klastrów dla zbioru Mall Customer Segmentation Data

Clusters	Silhouette
2	0.4907
3	0.4202
4	0.4468
5	0.4721
6	0.4943
7	0.5012
8	0.5109
9	0.5035
10	0.4843
11	0.4822
12	0.4709
13	0.4635
14	0.4602
15	0.4524
16	0.4415
17	0.4318
18	0.4317
19	0.4238
20	0.4106
21	0.4086
22	0.3945
23	0.3874
24	0.3851
25	0.3718
26	0.3784
27	0.3677
28	0.3641
29	0.3657
30	0.3554

Tabela 3. Wartość Silhouette w zależności od liczby klastrów dla zbioru Moons

Max iterations	Silhouette
1e2	0.6808
1e3	0.6808
1e4	0.6808
1e5	0.6808
1e6	0.6808
1e7	0.6808
1e8	0.6808
1e9	0.6808
1e10	0.6808
1e11	0.6808

Tabela 4. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 2 dla zbioru Iris

Max iterations	Silhouette
1e2	0.3682
1e3	0.3695
1e4	0.3672
1e5	0.3682
1e6	0.3634
1e7	0.3712
1e8	0.3634
1e9	0.3665
1e10	0.3682
1e11	0.3634

Tabela 5. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 6 dla zbioru Iris

Max iterations	Silhouette
1e2	0.3537
1e3	0.3631
1e4	0.3635
1e5	0.3464
1e6	0.3573
1e7	0.3615
1e8	0.3494
1e9	0.3631
1e10	0.3511
1e11	0.3415

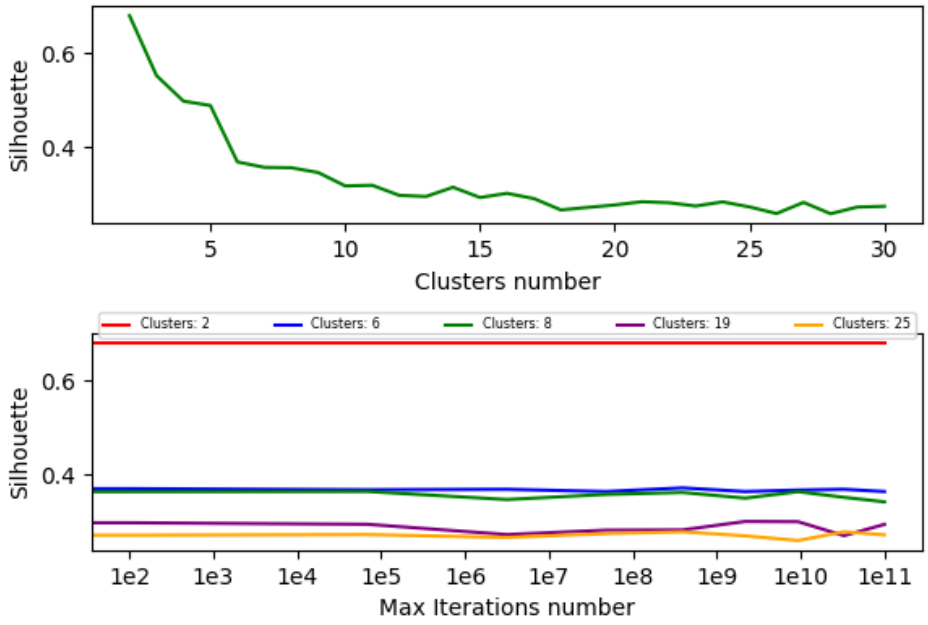
Tabela 6. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 8 dla zbioru Iris

Max iterations	Silhouette
1e2	0.2764
1e3	0.297
1e4	0.2937
1e5	0.2718
1e6	0.2812
1e7	0.2817
1e8	0.2997
1e9	0.2992
1e10	0.2695
1e11	0.2935

Tabela 7. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 19 dla zbioru Iris

Max iterations	Silhouette
1e2	0.2696
1e3	0.2705
1e4	0.2718
1e5	0.2657
1e6	0.2738
1e7	0.2777
1e8	0.2691
1e9	0.2592
1e10	0.2777
1e11	0.2712

Tabela 8. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 25 dla zbioru Iris



Rysunek 1. Wykresy wartości Silhouette w zależności od zmiennych dla zbioru Iris

Max iterations	Silhouette
1e2	0.2931
1e3	0.2931
1e4	0.2931
1e5	0.2931
1e6	0.2931
1e7	0.2931
1e8	0.2931
1e9	0.2931
1e10	0.2931
1e11	0.2931

Tabela 9. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 2 dla zbioru Mall Customer Segmentation Data

Max iterations	Silhouette
1e2	0.4521
1e3	0.4521
1e4	0.4521
1e5	0.4521
1e6	0.4521
1e7	0.4521
1e8	0.4521
1e9	0.4521
1e10	0.4521
1e11	0.4521

Tabela 10. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 6 dla zbioru Mall Customer Segmentation Data

Max iterations	Silhouette
1e2	0.426
1e3	0.4259
1e4	0.4259
1e5	0.4294
1e6	0.4278
1e7	0.4259
1e8	0.4275
1e9	0.4259
1e10	0.4295
1e11	0.4333

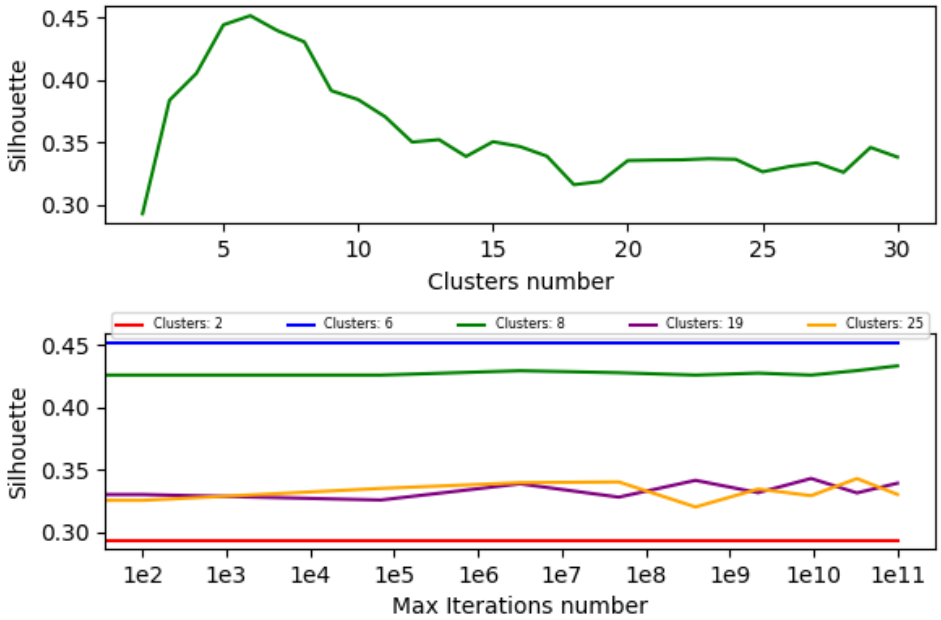
Tabela 11. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 8 dla zbioru Mall Customer Segmentation Data

Max iterations	Silhouette
1e2	0.346
1e3	0.33
1e4	0.3256
1e5	0.3386
1e6	0.3279
1e7	0.3414
1e8	0.3315
1e9	0.3429
1e10	0.3313
1e11	0.3391

Tabela 12. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 19 dla zbioru Mall Customer Segmentation Data

Max iterations	Silhouette
1e2	0.3388
1e3	0.3253
1e4	0.3349
1e5	0.3395
1e6	0.3401
1e7	0.32
1e8	0.3344
1e9	0.3291
1e10	0.3428
1e11	0.3301

Tabela 13. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 25 dla zbioru Mall Customer Segmentation Data



Rysunek 2. Wykresy wartości Silhouette w zależności od zmiennych dla zbioru Mall Customer Segmentation Data

Max iterations	Silhouette
1e2	0.4907
1e3	0.4907
1e4	0.4909
1e5	0.4907
1e6	0.4907
1e7	0.4907
1e8	0.4907
1e9	0.4907
1e10	0.4907
1e11	0.4907

Tabela 14. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 2 dla zbioru Moons

Max iterations	Silhouette
1e2	0.4945
1e3	0.4945
1e4	0.4938
1e5	0.494
1e6	0.4943
1e7	0.4945
1e8	0.4945
1e9	0.4943
1e10	0.4936
1e11	0.494

Tabela 15. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 6 dla zbioru Moons

Max iterations	Silhouette
1e2	0.5114
1e3	0.5114
1e4	0.5113
1e5	0.5139
1e6	0.5126
1e7	0.5114
1e8	0.5113
1e9	0.5114
1e10	0.5114
1e11	0.5114

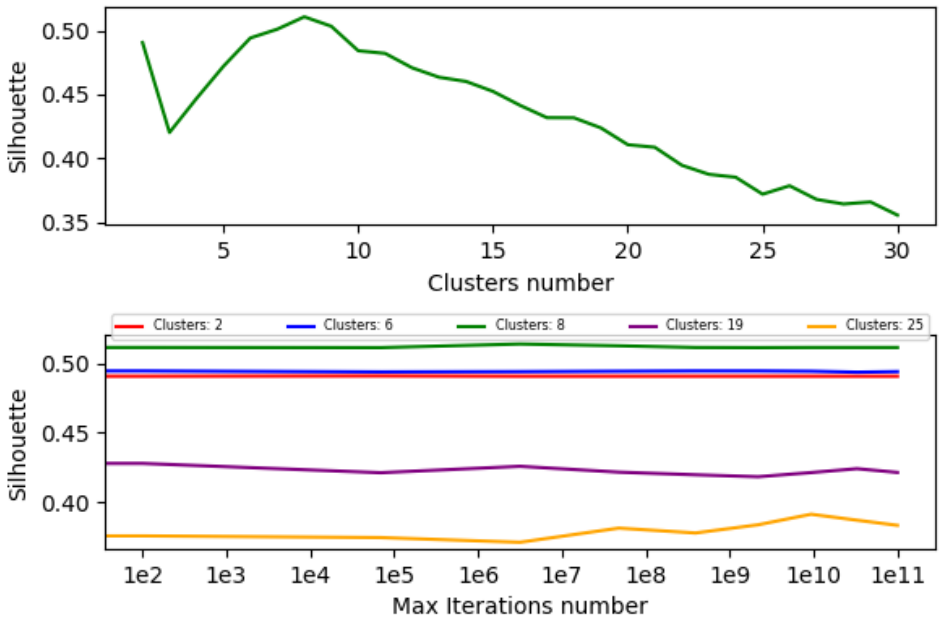
Tabela 16. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 8 dla zbioru Moons

Max iterations	Silhouette
1e2	0.4224
1e3	0.4279
1e4	0.4212
1e5	0.4258
1e6	0.4216
1e7	0.4197
1e8	0.4183
1e9	0.4213
1e10	0.4241
1e11	0.4214

Tabela 17. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 19 dla zbioru Moonss

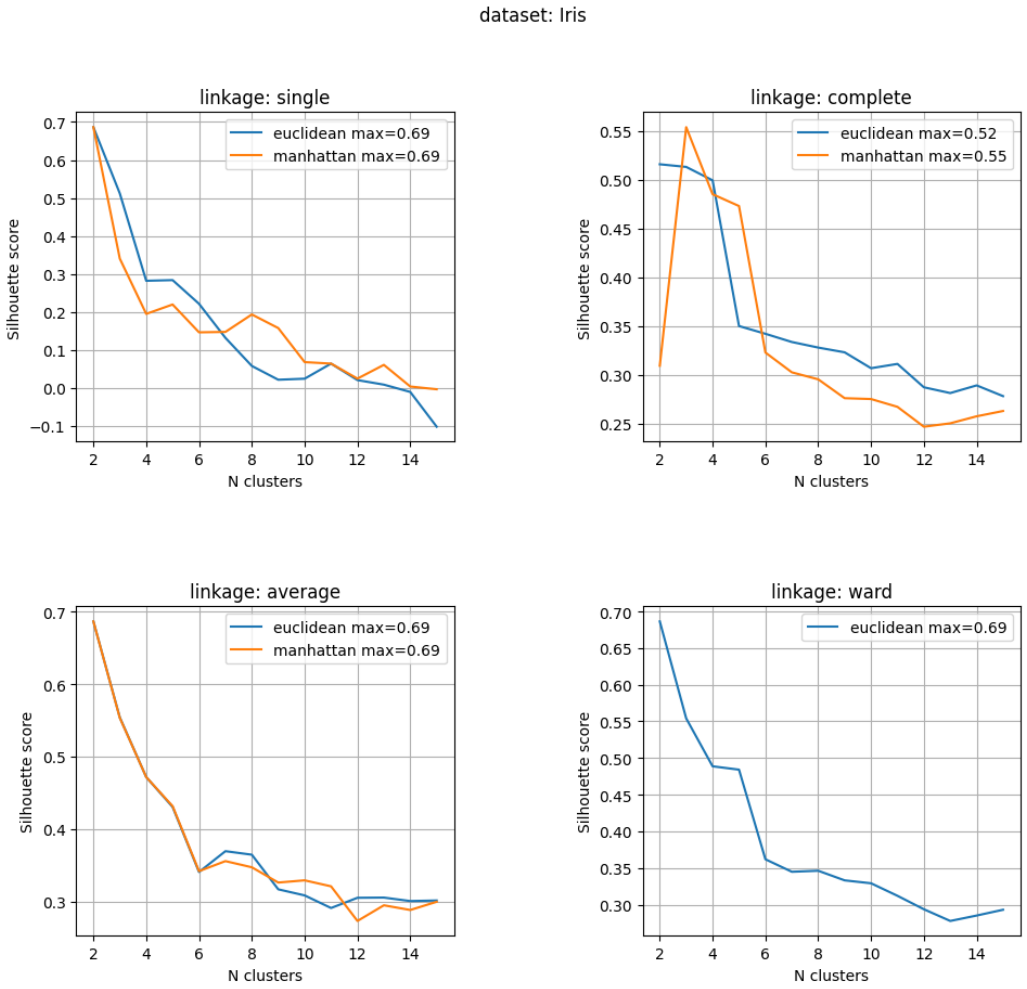
Max iterations	Silhouette
1e2	0.3757
1e3	0.3757
1e4	0.3745
1e5	0.3711
1e6	0.3813
1e7	0.3778
1e8	0.3837
1e9	0.3913
1e10	0.387
1e11	0.3833

Tabela 18. Wartość Silhouette w zależności od liczby maksymalnej iteracji dla liczby klastrów równej 25 dla zbioru Moons

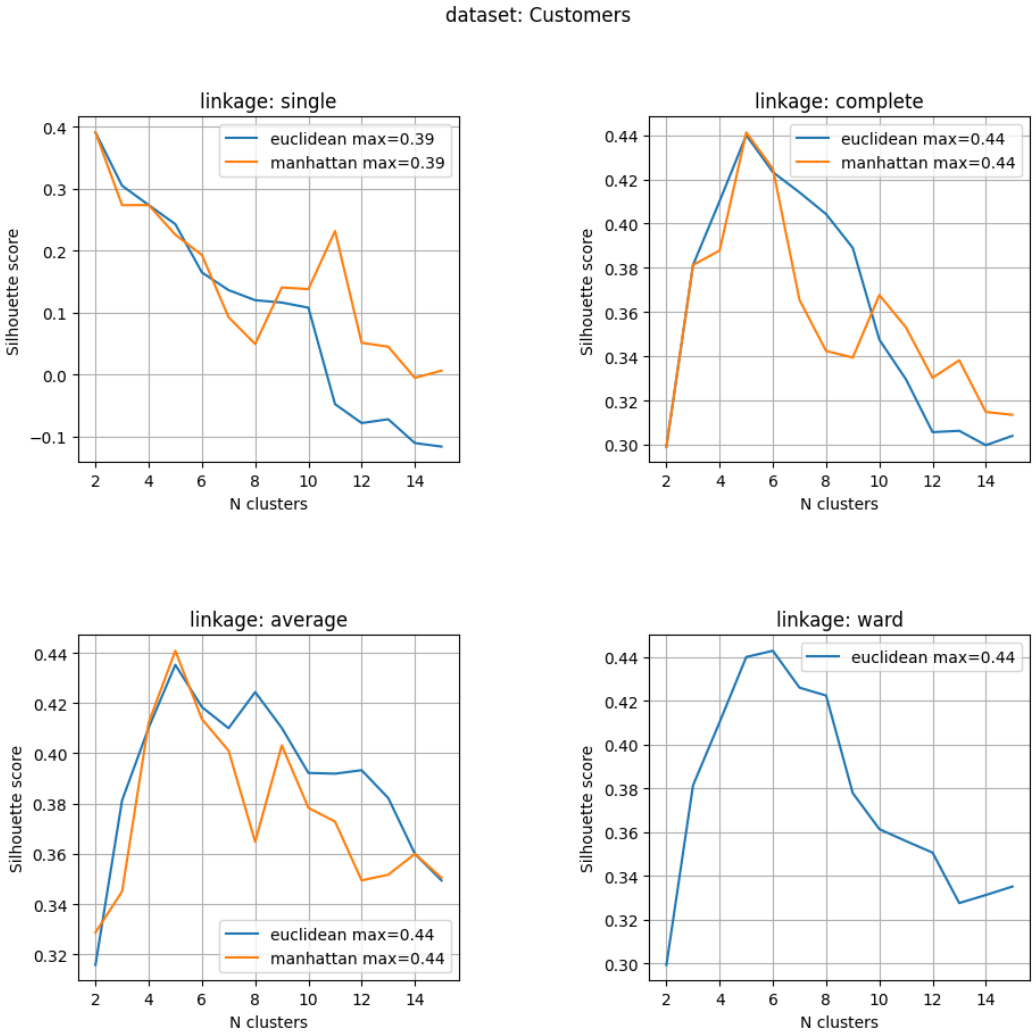


Rysunek 3. Wykresy wartości Silhouette w zależności od zmiennych dla zbioru Moons

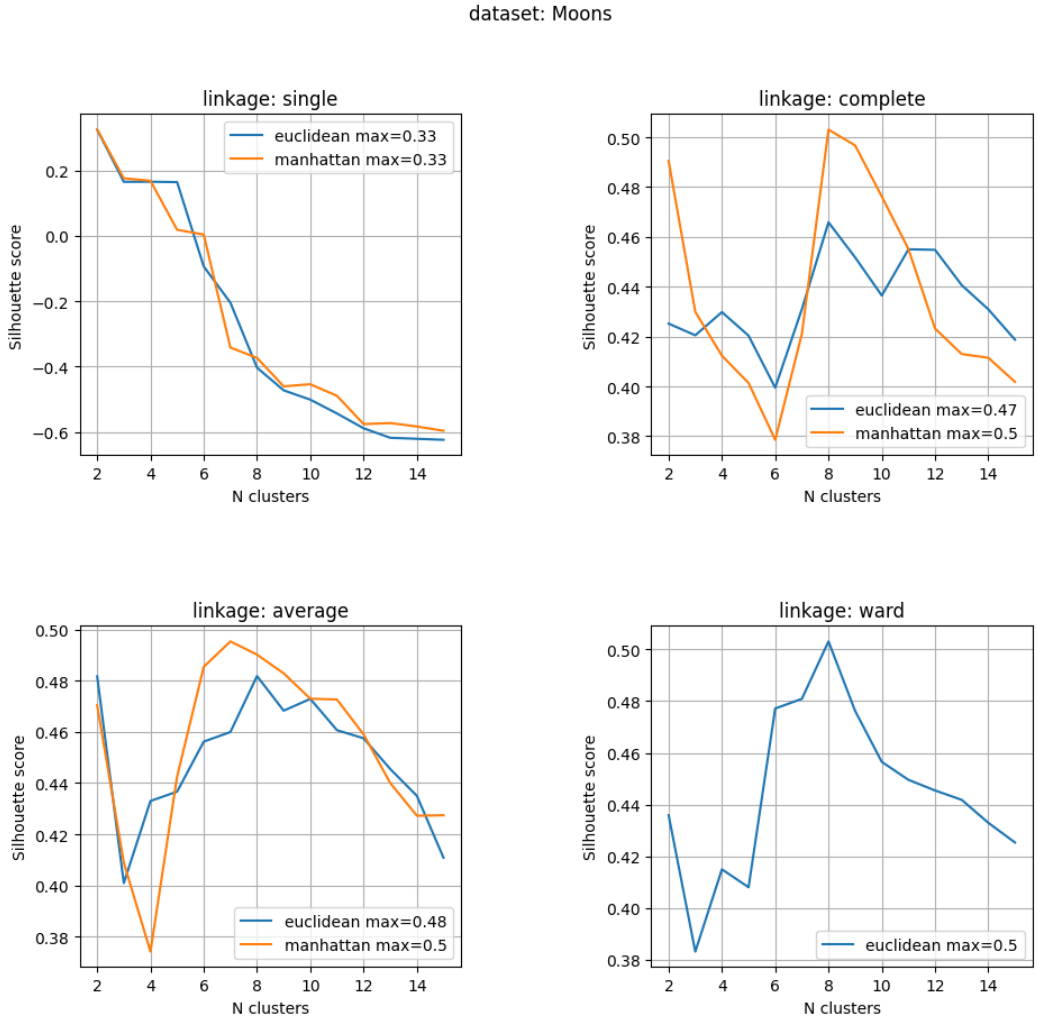
2.2. Algorytm aglomeracyjny



Rysunek 4. Wyniki działania algorytmu aglomeracyjnego dla zbioru danych , ,Iris Species”

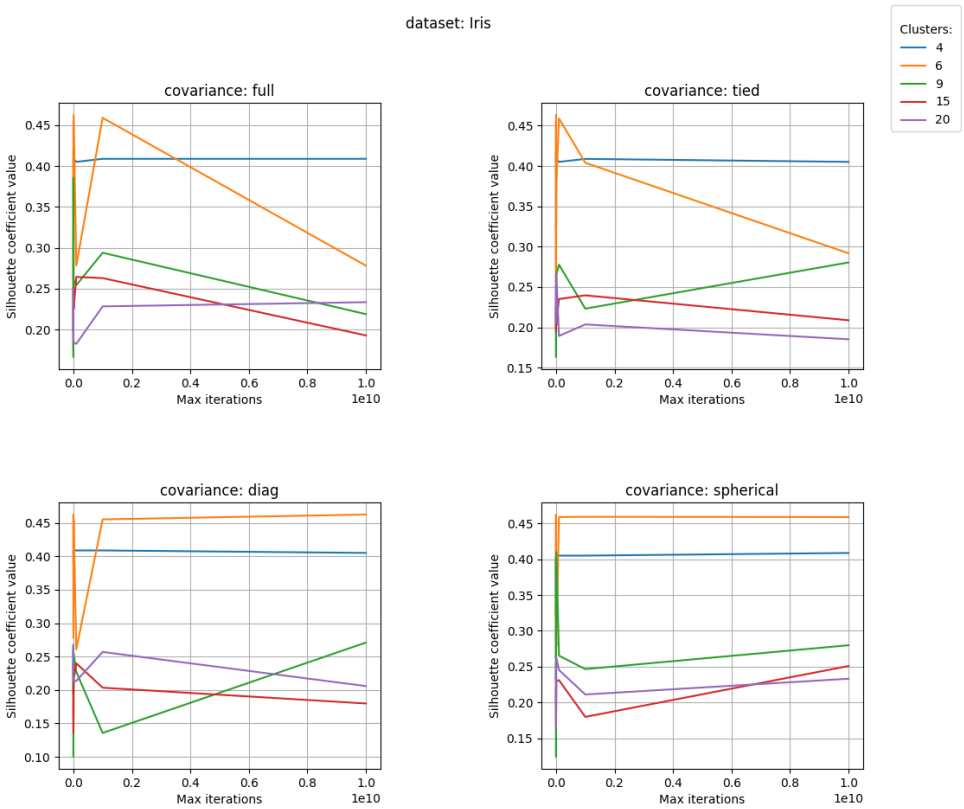


Rysunek 5. Wyniki działania algorytmu aglomeracyjnego dla zbioru danych , ,Mall Customer Segmentation Data”



Rysunek 6. Wyniki działania algorytmu aglomeracyjnego dla zbioru danych , ‚Moons”

2.3. Algorytm EM



Rysunek 7. Wyniki działania algorytmu EM dla zbioru danych Irysów w zależności od liczby iteracji i wybranej macierzy kowariancji dla różnej liczby klastrów

Max iterations	full	tied	diag	spherical
40	0.405	0.409	0.409	0.405
60	0.322	0.405	0.405	0.409
80	0.405	0.322	0.409	0.409
100	0.409	0.405	0.409	0.322
200	0.409	0.405	0.405	0.405
1000	0.409	0.405	0.405	0.376
10000	0.405	0.409	0.405	0.405
100000	0.405	0.418	0.409	0.322
1000000	0.409	0.322	0.405	0.409
10000000	0.409	0.409	0.409	0.405
100000000	0.405	0.405	0.409	0.405
1000000000	0.409	0.409	0.409	0.405
10000000000	0.409	0.405	0.405	0.409

Tabela 19. Wartości silhouette dla zbioru danych Irysów w zależności od liczby iteracji i wybranej macierzy kowariancji przy 4 klastrach

Max iterations	full	tied	diag	spherical
40	0.278	0.278	0.261	0.406
60	0.292	0.427	0.261	0.399
80	0.463	0.261	0.458	0.455
100	0.459	0.377	0.263	0.459
200	0.465	0.459	0.292	0.459
1000	0.28	0.261	0.292	0.261
10000	0.176	0.263	0.463	0.412
100000	0.406	0.456	0.459	0.463
1000000	0.261	0.263	0.389	0.292
10000000	0.463	0.377	0.455	0.278
100000000	0.278	0.459	0.261	0.459
1000000000	0.459	0.404	0.455	0.459
10000000000	0.278	0.292	0.463	0.459

Tabela 20. Wartości silhouette dla zbioru danych Irysów w zależności od liczby iteracji i wybranej macierzy kowariancji przy 6 klastrach

Max iterations	full	tied	diag	spherical
40	0.241	0.394	0.243	0.246
60	0.257	0.435	0.163	0.147
80	0.243	0.254	0.278	0.27
100	0.216	0.248	0.248	0.229
200	0.245	0.211	0.162	0.277
1000	0.244	0.203	0.26	0.124
10000	0.23	0.226	0.196	0.401
100000	0.166	0.163	0.268	0.233
1000000	0.384	0.227	0.1	0.28
10000000	0.26	0.265	0.255	0.41
100000000	0.254	0.278	0.227	0.265
1000000000	0.294	0.223	0.136	0.247
10000000000	0.219	0.28	0.271	0.28

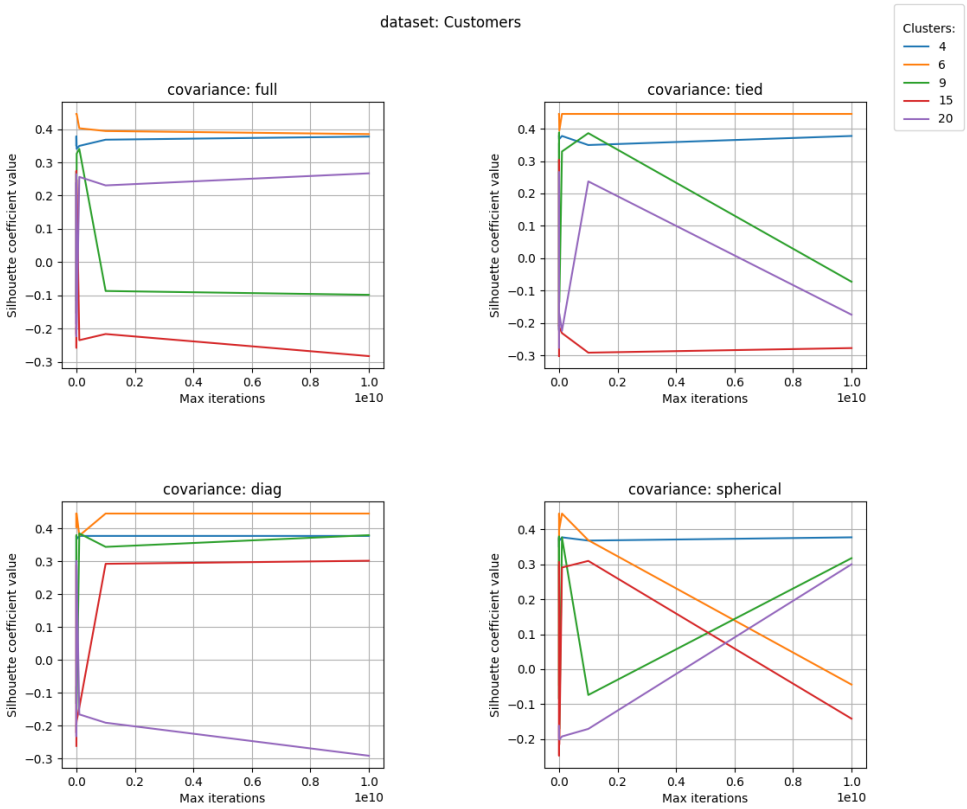
Tabela 21. Wartości silhouette dla zbioru danych Irysów w zależności od liczby iteracji i wybranej macierzy kowariancji przy 9 klastrach

Max iterations	full	tied	diag	spherical
40	0.202	0.182	0.159	0.247
60	0.246	0.173	0.196	0.241
80	0.168	0.175	0.227	0.227
100	0.197	0.233	0.202	0.18
200	0.236	0.138	0.147	0.24
1000	0.243	0.218	0.201	0.212
10000	0.198	0.203	0.19	0.218
100000	0.211	0.196	0.155	0.219
1000000	0.245	0.217	0.136	0.176
10000000	0.225	0.202	0.217	0.23
100000000	0.264	0.235	0.24	0.231
1000000000	0.263	0.24	0.203	0.18
10000000000	0.193	0.209	0.18	0.251

Tabela 22. Wartości silhouette dla zbioru danych Irysów w zależności od liczby iteracji i wybranej macierzy kowariancji przy 15 klastrach

Max iterations	full	tied	diag	spherical
40	0.204	0.227	0.232	0.235
60	0.219	0.261	0.231	0.095
80	0.239	0.238	0.253	0.232
100	0.182	0.26	0.222	0.194
200	0.195	0.238	0.187	0.227
1000	0.226	0.25	0.24	0.235
10000	0.198	0.206	0.219	0.231
100000	0.249	0.262	0.267	0.244
1000000	0.185	0.205	0.228	0.165
10000000	0.184	0.267	0.213	0.263
100000000	0.182	0.189	0.213	0.245
1000000000	0.228	0.204	0.257	0.211
10000000000	0.233	0.185	0.206	0.233

Tabela 23. Wartości silhouette dla zbioru danych Irysów w zależności od liczby iteracji i wybranej macierzy kowariancji przy 20 klastrach



Rysunek 8. Wyniki działania algorytmu EM dla zbioru danych Customers w zależności od liczby iteracji i wybranej macierzy kowariancji dla różnej liczby klastrów

Max iterations	full	tied	diag	spherical
40	0.368	0.377	0.377	0.377
60	0.377	0.377	0.377	0.377
80	0.377	0.368	0.377	0.377
100	0.377	0.377	0.377	0.377
200	0.377	0.368	0.377	0.377
1000	0.349	0.377	0.377	0.377
10000	0.377	0.377	0.377	0.349
100000	0.377	0.377	0.377	0.377
1000000	0.377	0.368	0.377	0.368
10000000	0.34	0.368	0.368	0.368
100000000	0.349	0.377	0.377	0.377
1000000000	0.368	0.349	0.377	0.368
10000000000	0.377	0.377	0.377	0.377

Tabela 24. Wartości silhouette dla zbioru danych Customers w zależności od liczby iteracji i wybranej macierzy kowariancji przy 4 klastrach

Max iterations	full	tied	diag	spherical
40	0.371	0.446	0.446	-0.006
60	0.446	0.394	0.383	0.394
80	0.394	0.001	0.446	0.446
100	0.406	0.446	0.446	0.402
200	0.385	0.446	0.446	0.446
1000	0.446	0.446	0.446	0.446
10000	0.446	0.446	0.403	0.446
100000	0.446	0.446	0.446	0.446
1000000	0.446	0.446	0.446	0.446
10000000	0.446	0.394	0.446	0.402
100000000	0.402	0.446	0.378	0.446
1000000000	0.394	0.446	0.446	0.369
10000000000	0.385	0.446	0.446	-0.044

Tabela 25. Wartości silhouette dla zbioru danych Customers w zależności od liczby iteracji i wybranej macierzy kowariancji przy 6 klastrach

Max iterations	full	tied	diag	spherical
40	-0.073	0.363	0.387	0.28
60	-0.137	-0.086	-0.086	0.379
80	-0.136	-0.232	-0.133	0.346
100	-0.074	0.353	0.361	0.292
200	0.354	-0.299	-0.074	-0.112
1000	0.273	0.349	-0.094	-0.085
10000	-0.153	-0.22	-0.133	0.374
100000	-0.113	-0.071	0.305	0.328
1000000	-0.086	0.388	0.38	0.294
10000000	0.326	-0.135	-0.156	-0.17
100000000	0.34	0.33	0.385	0.377
1000000000	-0.086	0.386	0.344	-0.074
10000000000	-0.098	-0.073	0.38	0.318

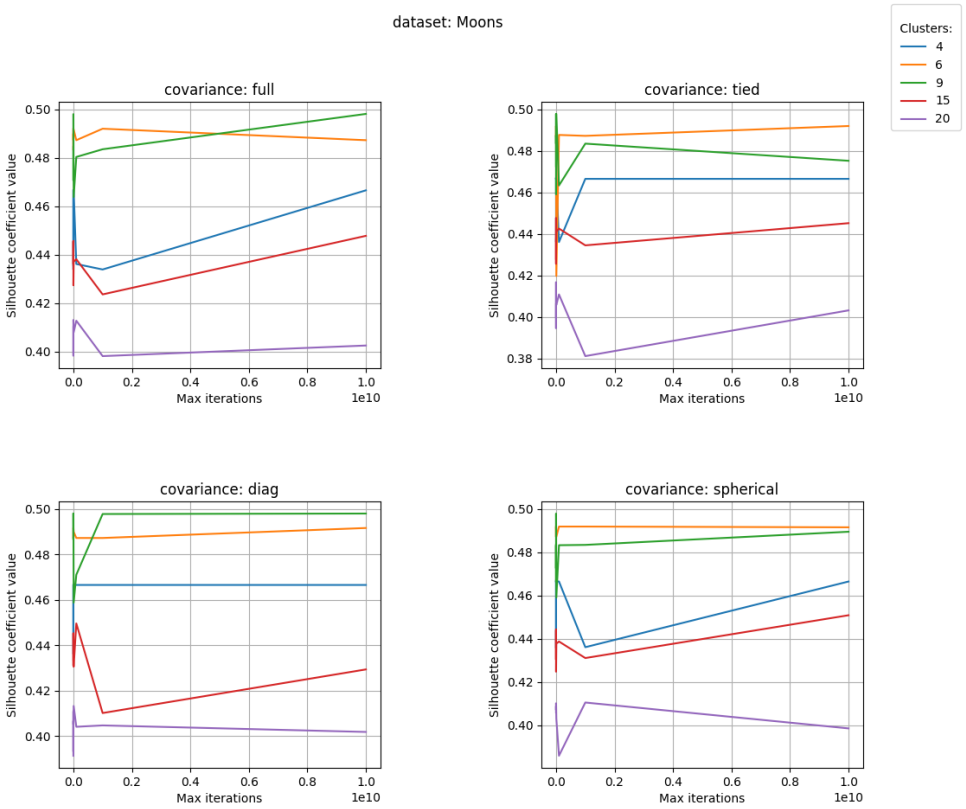
Tabela 26. Wartości silhouette dla zbioru danych Customers w zależności od liczby iteracji i wybranej macierzy kowariancji przy 9 klastrach

Max iterations	full	tied	diag	spherical
40	0.327	0.273	0.319	0.27
60	-0.167	-0.204	0.281	0.303
80	-0.275	-0.181	0.291	-0.246
100	0.326	-0.198	0.292	0.258
200	0.247	0.301	-0.274	-0.292
1000	-0.257	0.304	0.283	0.302
10000	-0.163	0.299	0.29	-0.248
100000	0.275	-0.21	-0.221	-0.197
1000000	-0.216	-0.206	-0.195	0.299
10000000	0.248	-0.214	-0.185	-0.215
100000000	-0.234	-0.231	-0.144	0.291
1000000000	-0.216	-0.291	0.293	0.31
10000000000	-0.282	-0.277	0.302	-0.142

Tabela 27. Wartości silhouette dla zbioru danych Customers w zależności od liczby iteracji i wybranej macierzy kowariancji przy 15 klastrach

Max iterations	full	tied	diag	spherical
40	-0.13	-0.187	-0.223	-0.212
60	-0.259	-0.2	-0.225	-0.167
80	-0.235	-0.202	-0.151	-0.22
100	0.295	-0.227	0.301	0.314
200	0.279	-0.202	0.285	-0.254
1000	-0.186	-0.184	-0.232	-0.182
10000	0.265	0.267	-0.18	-0.165
100000	-0.223	-0.214	0.256	-0.161
1000000	0.239	-0.223	0.297	-0.198
10000000	-0.202	-0.171	0.281	-0.201
100000000	0.256	-0.224	-0.165	-0.192
1000000000	0.231	0.237	-0.191	-0.171
10000000000	0.267	-0.174	-0.291	0.3

Tabela 28. Wartości silhouette dla zbioru danych Customers w zależności od liczby iteracji i wybranej macierzy kowariancji przy 20 klastrach



Rysunek 9. Wyniki działania algorytmu EM dla zbioru danych Moons w zależności od liczby iteracji i wybranej macierzy kowariancji dla różnej liczby klastrów

Max iterations	full	tied	diag	spherical
40	0.467	0.436	0.436	0.467
60	0.467	0.436	0.467	0.467
80	0.467	0.467	0.436	0.467
100	0.467	0.467	0.467	0.434
200	0.467	0.467	0.467	0.467
1000	0.434	0.467	0.436	0.467
10000	0.436	0.467	0.467	0.467
100000	0.436	0.467	0.467	0.436
1000000	0.467	0.436	0.467	0.467
10000000	0.467	0.467	0.467	0.467
100000000	0.436	0.436	0.467	0.467
1000000000	0.434	0.467	0.467	0.436
10000000000	0.467	0.467	0.467	0.467

Tabela 29. Wartości silhouette dla zbioru danych Moons w zależności od liczby iteracji i wybranej macierzy kowariancji przy 4 klastrach

Max iterations	full	tied	diag	spherical
40	0.488	0.492	0.492	0.409
60	0.492	0.492	0.487	0.487
80	0.488	0.487	0.487	0.492
100	0.492	0.492	0.492	0.487
200	0.492	0.492	0.487	0.492
1000	0.483	0.492	0.487	0.492
10000	0.49	0.492	0.492	0.487
100000	0.487	0.487	0.492	0.487
1000000	0.492	0.488	0.492	0.492
10000000	0.492	0.42	0.49	0.487
100000000	0.487	0.488	0.487	0.492
1000000000	0.492	0.487	0.487	0.492
10000000000	0.487	0.492	0.492	0.492

Tabela 30. Wartości silhouette dla zbioru danych Moons w zależności od liczby iteracji i wybranej macierzy kowariancji przy 6 klastrach

Max iterations	full	tied	diag	spherical
40	0.463	0.498	0.487	0.49
60	0.463	0.488	0.498	0.498
80	0.498	0.473	0.459	0.498
100	0.498	0.481	0.498	0.498
200	0.498	0.458	0.471	0.486
1000	0.483	0.487	0.498	0.473
10000	0.471	0.483	0.486	0.483
100000	0.471	0.498	0.498	0.466
1000000	0.498	0.459	0.498	0.498
10000000	0.464	0.498	0.459	0.459
100000000	0.48	0.463	0.471	0.483
1000000000	0.483	0.483	0.498	0.483
10000000000	0.498	0.475	0.498	0.49

Tabela 31. Wartości silhouette dla zbioru danych Moons w zależności od liczby iteracji i wybranej macierzy kowariancji przy 9 klastrach

Max iterations	full	tied	diag	spherical
40	0.423	0.44	0.448	0.435
60	0.451	0.451	0.422	0.434
80	0.446	0.436	0.438	0.423
100	0.443	0.448	0.433	0.439
200	0.44	0.421	0.42	0.436
1000	0.445	0.448	0.435	0.43
10000	0.436	0.446	0.431	0.444
100000	0.446	0.426	0.445	0.429
1000000	0.427	0.437	0.434	0.425
10000000	0.437	0.441	0.431	0.438
100000000	0.438	0.443	0.45	0.439
1000000000	0.424	0.435	0.41	0.431
10000000000	0.448	0.445	0.429	0.451

Tabela 32. Wartości silhouette dla zbioru danych Moons w zależności od liczby iteracji i wybranej macierzy kowariancji przy 15 klastrach

Max iterations	full	tied	diag	spherical
40	0.415	0.391	0.401	0.396
60	0.41	0.384	0.411	0.41
80	0.394	0.412	0.391	0.395
100	0.396	0.389	0.393	0.412
200	0.397	0.413	0.403	0.404
1000	0.4	0.417	0.407	0.409
10000	0.398	0.395	0.393	0.409
100000	0.402	0.401	0.411	0.407
1000000	0.413	0.406	0.405	0.41
10000000	0.408	0.406	0.413	0.403
100000000	0.413	0.411	0.404	0.386
1000000000	0.398	0.381	0.405	0.411
10000000000	0.402	0.403	0.402	0.399

Tabela 33. Wartości silhouette dla zbioru danych Moons w zależności od liczby iteracji i wybranej macierzy kowariancji przy 20 klastrach

2.4. Algorytm DBSCAN

2.4.1. Metryka Euklidesowa

Eps	Silh
0.1	-0.5147
0.2	-0.2479
0.3	-0.0026
0.4	0.2419
0.5	0.1904
0.6	0.3646
0.7	0.2818
0.8	0.5118
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

Tabela 34. Iris, min samples: 2

Eps	Silh
0.1	0.0439
0.2	-0.3396
0.3	0.0319
0.4	0.3346
0.5	0.3463
0.6	0.4226
0.7	0.5016
0.8	0.5118
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

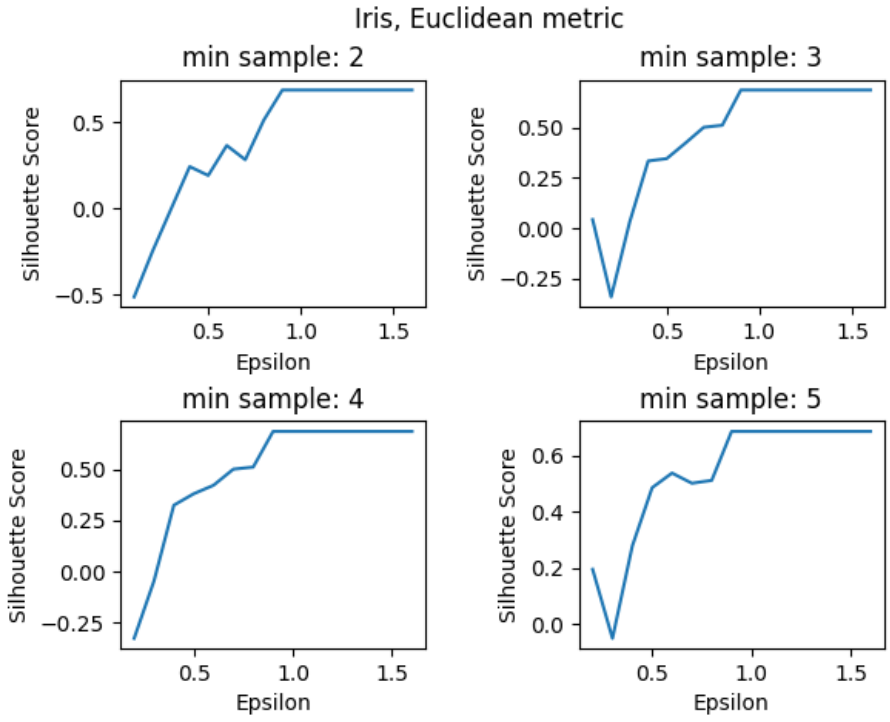
Tabela 35. Iris, min samples: 3

Eps	Silh
0.2	-0.3275
0.3	-0.0463
0.4	0.3249
0.5	0.3809
0.6	0.4226
0.7	0.5016
0.8	0.5118
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

Tabela 36. Iris, min samples: 4

Eps	Silh
0.2	0.1945
0.3	-0.0518
0.4	0.2778
0.5	0.4858
0.6	0.5379
0.7	0.5016
0.8	0.5118
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

Tabela 37. Iris, min samples: 5



Rysunek 10.

Eps	Silh
0.2	0.1852
0.3	0.0408
0.4	0.2778
0.5	0.477
0.6	0.5454
0.7	0.5251
0.8	0.5118
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

Tabela 38. Iris, min samples: 6

Eps	Silh
0.2	0.1734
0.3	0.0085
0.4	0.1992
0.5	0.4641
0.6	0.5419
0.7	0.5346
0.8	0.522
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

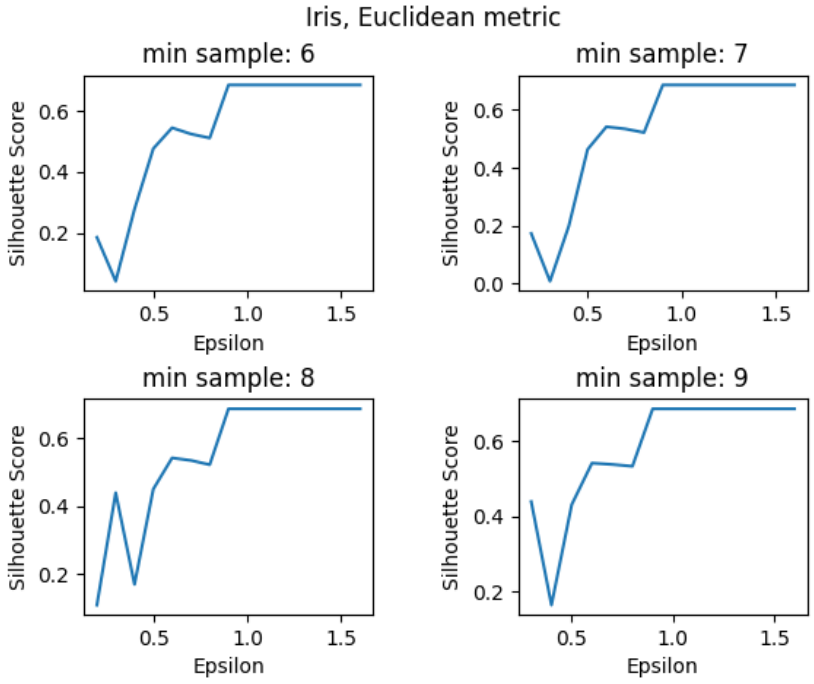
Tabela 39. Iris, min samples: 7

Eps	Silh
0.2	0.1092
0.3	0.4396
0.4	0.17
0.5	0.4502
0.6	0.5419
0.7	0.5346
0.8	0.522
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

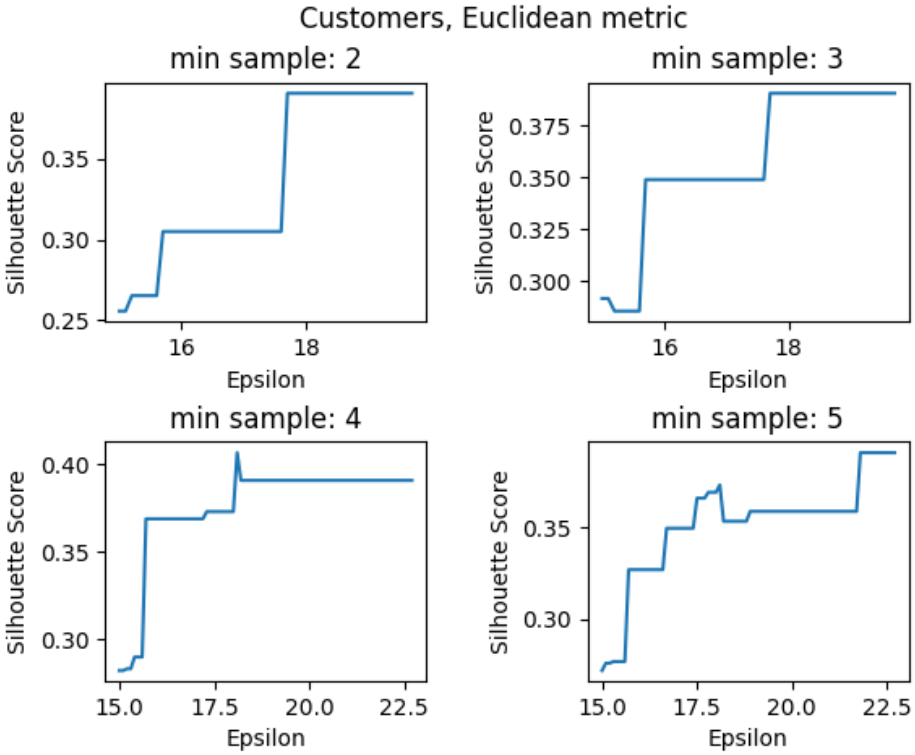
Tabela 40. Iris, min samples: 8

Eps	Silh
0.3	0.4396
0.4	0.1643
0.5	0.4314
0.6	0.5419
0.7	0.5384
0.8	0.5336
0.9	0.6864
1.0	0.6864
1.1	0.6864
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864

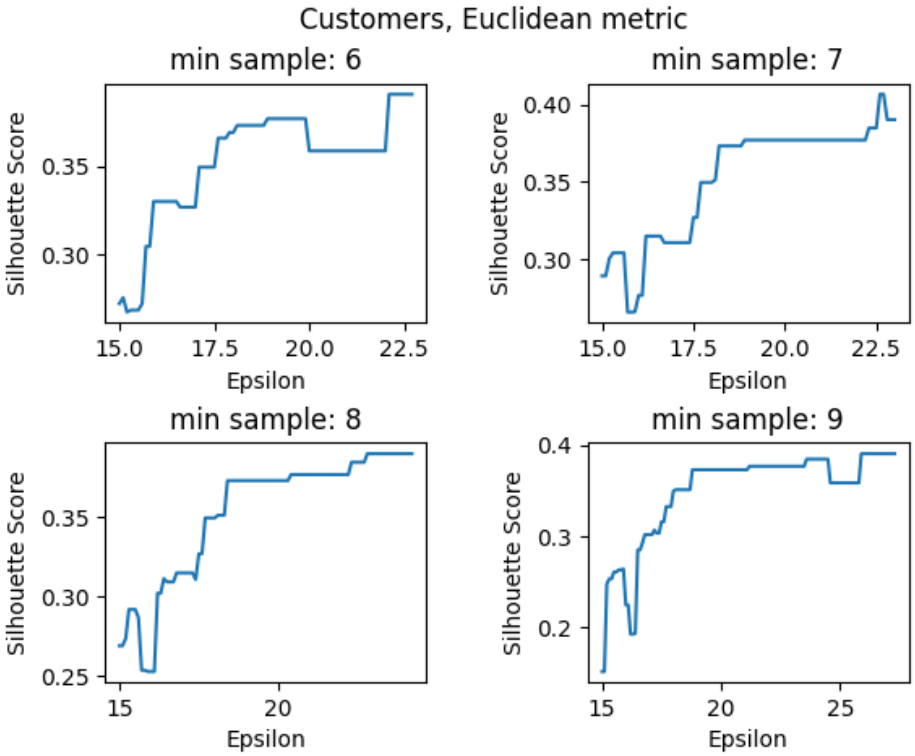
Tabela 41. Iris, min samples: 9



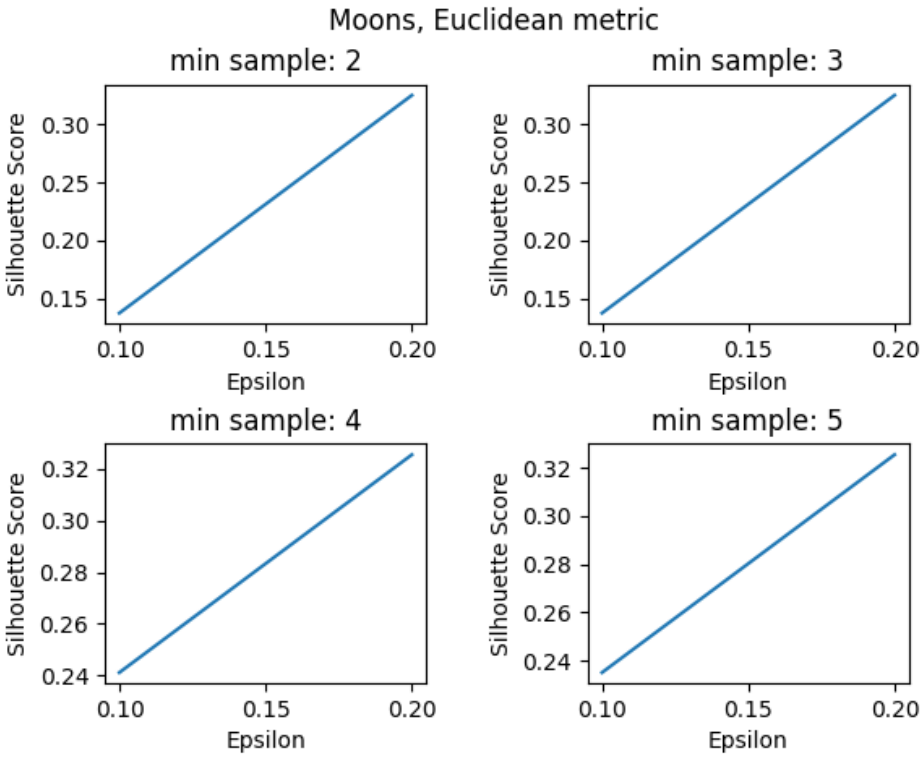
Rysunek 11.



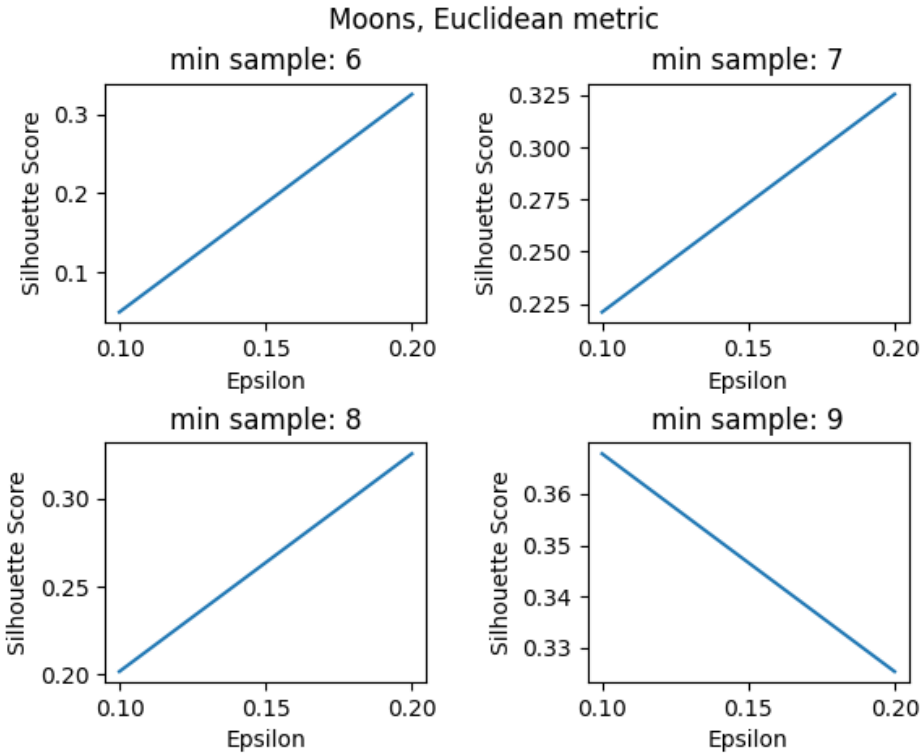
Rysunek 12.



Rysunek 13.



Rysunek 14.



Rysunek 15.

Eps	Silh
0.1	0.1371
0.2	0.3254

Tabela 42. Moons, min samples: 2

Eps	Silh
0.1	0.1371
0.2	0.3254

Tabela 43. Moons, min samples: 3

Eps	Silh
0.1	0.2412
0.2	0.3254

Tabela 44. Moons, min samples: 4

Eps	Silh
0.1	0.235
0.2	0.3254

Tabela 45. Moons, min samples: 5

Eps	Silh
0.1	0.0492
0.2	0.3254

Tabela 46. Moons, min samples: 6

Eps	Silh
0.1	0.2211
0.2	0.3254

Tabela 47. Moons, min samples: 7

Eps	Silh
0.1	0.2018
0.2	0.3254

Tabela 48. Moons, min samples: 8

Eps	Silh
0.1	0.3678
0.2	0.3254

Tabela 49. Moons, min samples: 9

2.4.2. Metryka Manhattan

Eps	Silh
0.1	-0.5147
0.2	-0.4523
0.3	-0.2502
0.4	-0.0944
0.5	0.002
0.6	0.2219
0.7	0.1472
0.8	0.372
0.9	0.2968
1.0	0.4082
1.1	0.4082
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

Tabela 50. Iris, min samples: 2

Eps	Silh
0.1	0.0439
0.2	0.0758
0.3	-0.3269
0.4	-0.069
0.5	0.0106
0.6	0.294
0.7	0.2805
0.8	0.3671
0.9	0.3491
1.0	0.4873
1.1	0.4873
1.2	0.6864
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

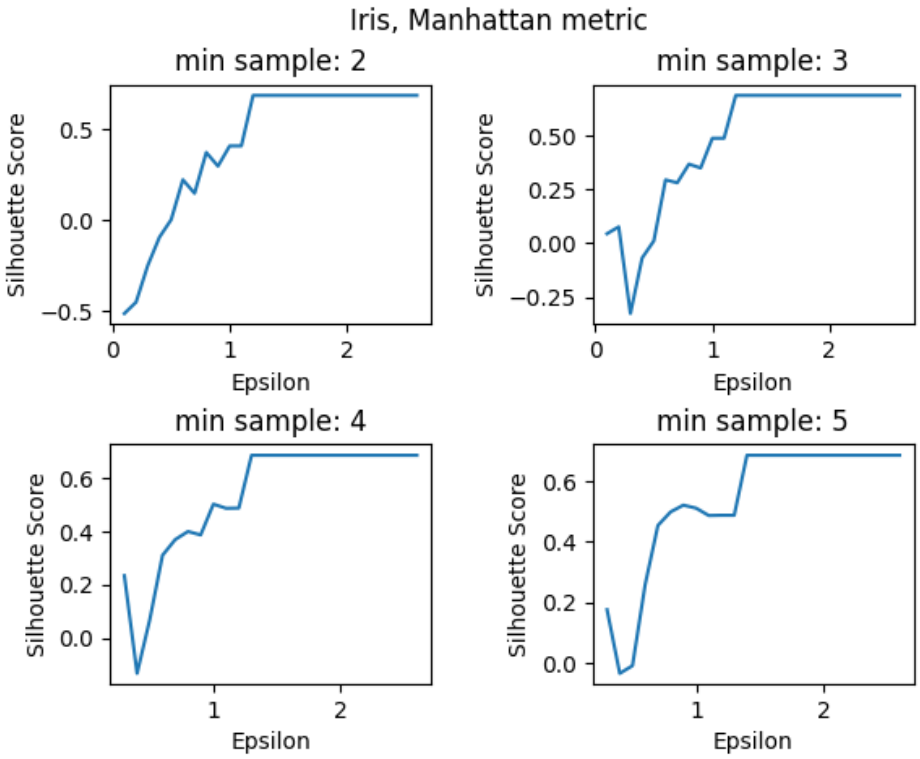
Tabela 51. Iris, min samples: 3

Eps	Silh
0.3	0.2341
0.4	-0.133
0.5	0.0696
0.6	0.3107
0.7	0.3698
0.8	0.4001
0.9	0.3872
1.0	0.5028
1.1	0.4873
1.2	0.4881
1.3	0.6864
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

Tabela 52. Iris, min samples: 4

Eps	Silh
0.3	0.1762
0.4	-0.0344
0.5	-0.009
0.6	0.2596
0.7	0.4545
0.8	0.4994
0.9	0.5216
1.0	0.5117
1.1	0.4873
1.2	0.4881
1.3	0.4881
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

Tabela 53. Iris, min samples: 5


Iris, Manhattan metric
min sample: 2
min sample: 3
min sample: 4
min sample: 5
Silhouette Score
Epsilon

Rysunek 16.

Eps	Silh
0.3	0.1232
0.4	-0.1083
0.5	-0.0258
0.6	0.1805
0.7	0.4414
0.8	0.498
0.9	0.5211
1.0	0.5298
1.1	0.5097
1.2	0.4881
1.3	0.4881
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

Tabela 54. Iris, min samples: 6

Eps	Silh
0.3	0.1232
0.4	0.3256
0.5	0.1163
0.6	0.124
0.7	0.4152
0.8	0.498
0.9	0.5211
1.0	0.533
1.1	0.5193
1.2	0.5446
1.3	0.4881
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

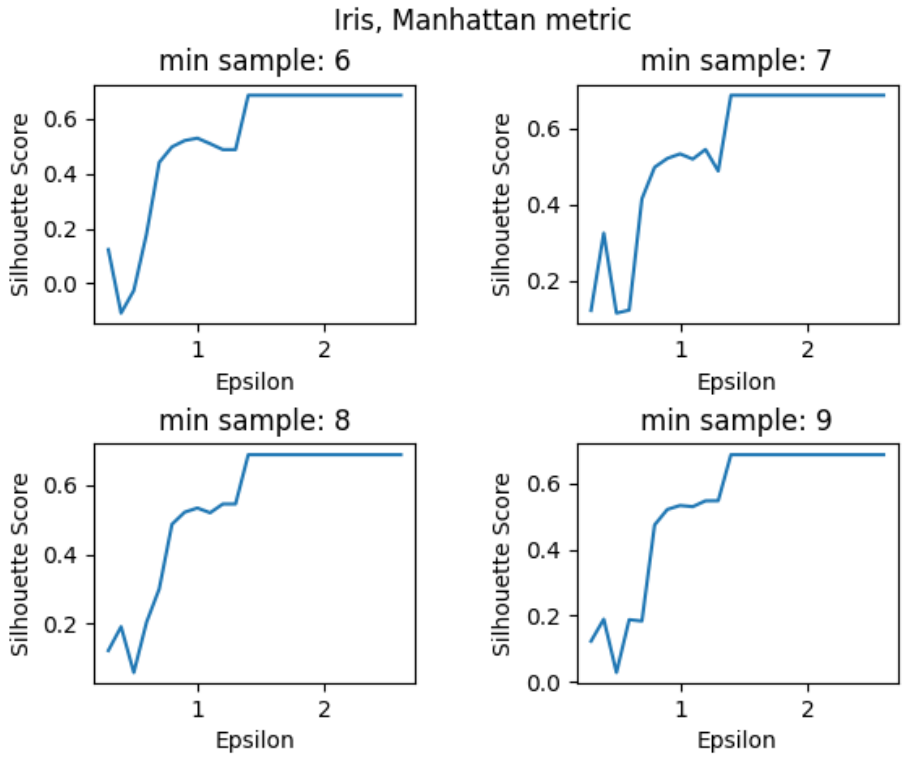
Tabela 55. Iris, min samples: 7

Eps	Silh
0.3	0.1232
0.4	0.1917
0.5	0.0599
0.6	0.2044
0.7	0.3005
0.8	0.4862
0.9	0.5211
1.0	0.533
1.1	0.5193
1.2	0.5446
1.3	0.5446
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

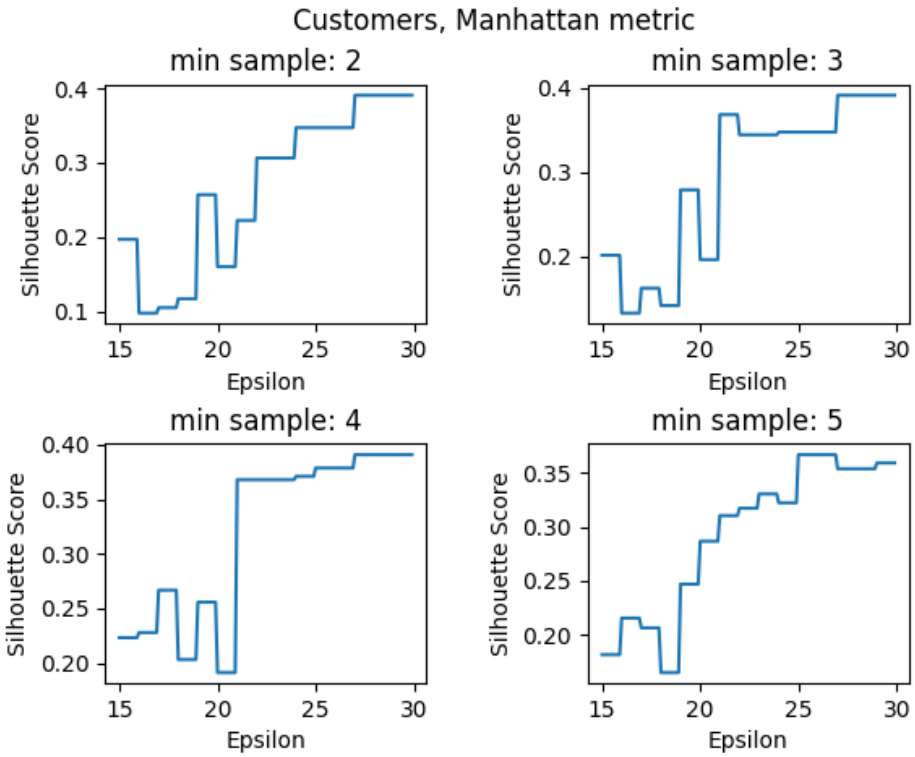
Tabela 56. Iris, min samples: 8

Eps	Silh
0.3	0.1232
0.4	0.1897
0.5	0.0286
0.6	0.1882
0.7	0.1841
0.8	0.4744
0.9	0.5211
1.0	0.533
1.1	0.5295
1.2	0.5471
1.3	0.5474
1.4	0.6864
1.5	0.6864
1.6	0.6864
1.7	0.6864
1.8	0.6864
1.9	0.6864
2.0	0.6864
2.1	0.6864
2.2	0.6864
2.3	0.6864
2.4	0.6864
2.5	0.6864
2.6	0.6864

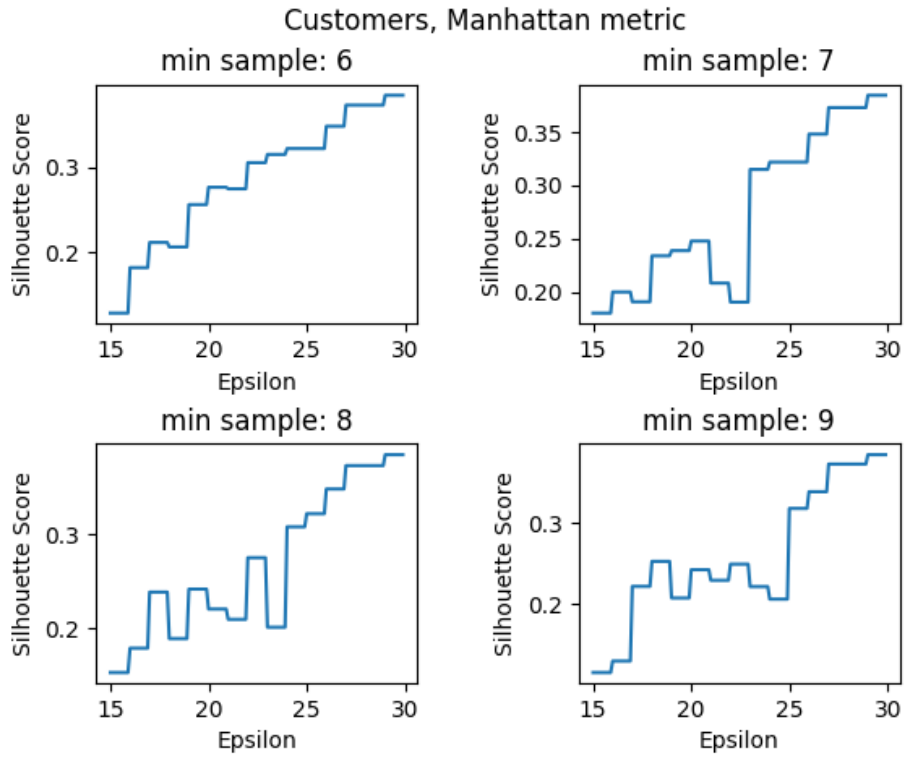
Tabela 57. Iris, min samples: 9



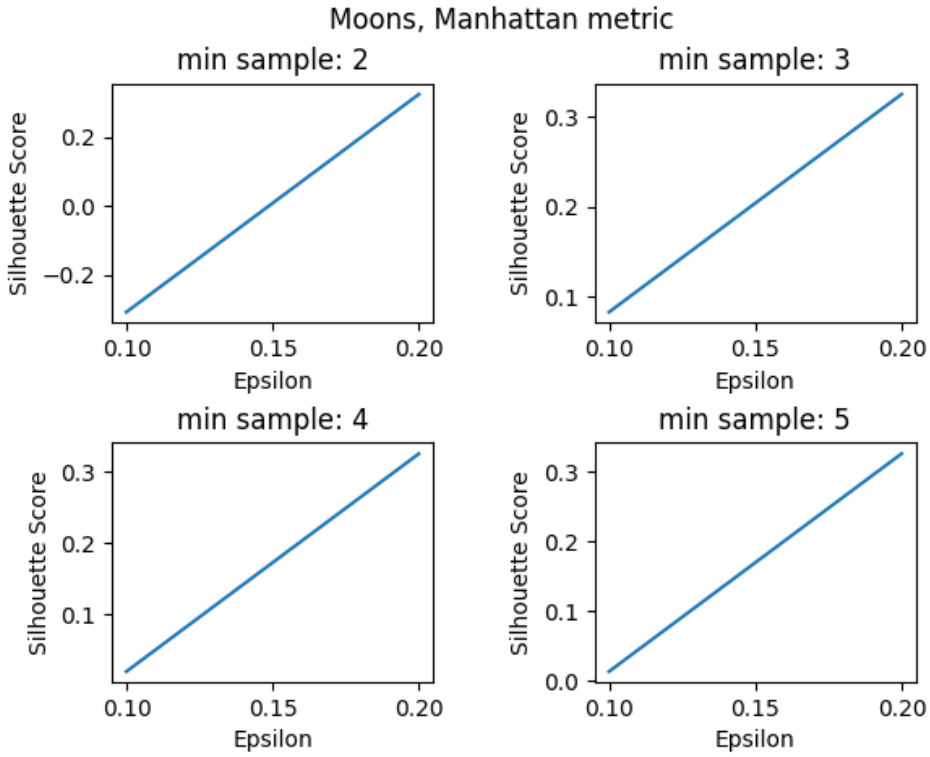
Rysunek 17.



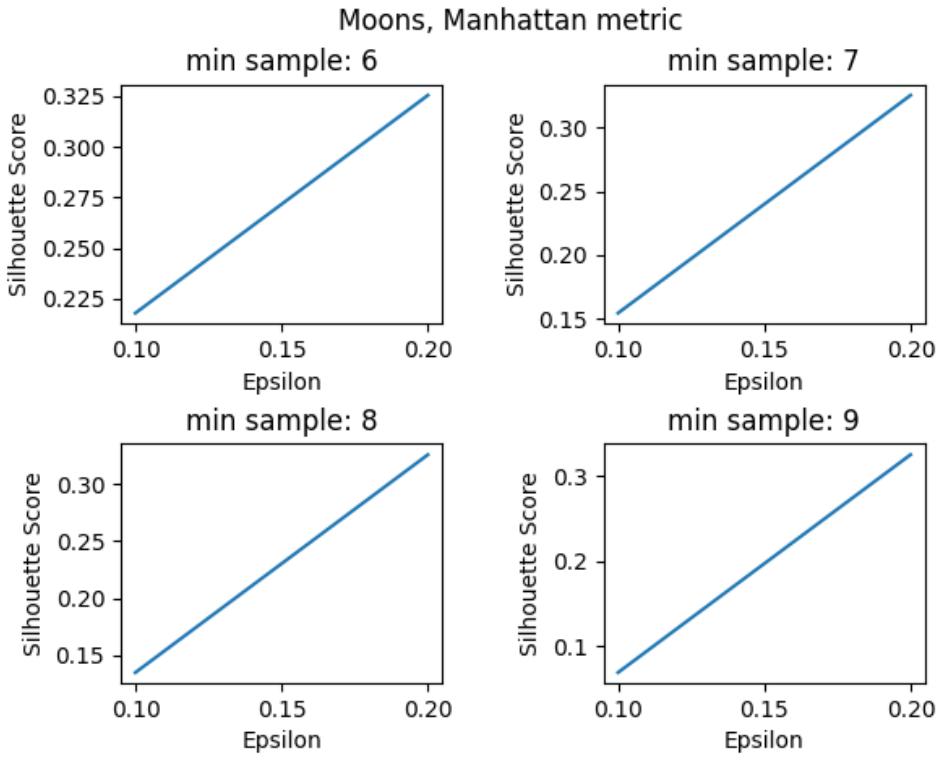
Rysunek 18.



Rysunek 19.



Rysunek 20.



Rysunek 21.

Eps	Silh
0.1	-0.3083
0.2	0.3254

Tabela 58. Moons, min samples: 2

Eps	Silh
0.1	0.083
0.2	0.3254

Tabela 59. Moons, min samples: 3

Eps	Silh
0.1	0.0192
0.2	0.3254

Tabela 60. Moons, min samples: 4

Eps	Silh
0.1	0.0136
0.2	0.3254

Tabela 61. Moons, min samples: 5

Eps	Silh
0.1	0.218
0.2	0.3254

Tabela 62. Moons, min samples: 6

Eps	Silh
0.1	0.1543
0.2	0.3254

Tabela 63. Moons, min samples: 7

Eps	Silh
0.1	0.135
0.2	0.3254

Tabela 64. Moons, min samples: 8

Eps	Silh
0.1	0.0688
0.2	0.3254

Tabela 65. Moons, min samples: 9

3. Dyskusja

3.1. Algorytm k-średnich

Analizując wartości współczynnika zarysu dla zbioru [1], dzięki tabeli 1 oraz wizualizacji tych danych zawartych na wykresie 1 można wywnioskować, że optymalną liczbą klastrów jest liczba 2, natomiast wartości współczynnika

zarysu dla liczby klastrów większej od 5 drastycznie spadają, przez co należy uznać je jako nieodpowiednie dla tego zbioru. Natomiast badanie wpływu wartości maksymalnej liczby iteracji wskazało, że największy wpływ na wartość współczynnika zarysu ma liczba z przedziału [1e9; 1e11] dla liczby klastrów równej 19 lub 25. Niestety dla pozostałych wartości liczby klastrów nie zauważono żadnej wartości maksymalnej liczby iteracji, którą należałoby dokładniej zbadać - nie udało się ustalić rzędu wielkości tego parametru, który mógłby mieć wpływ na końcowy wynik.

Analizując wartości współczynnika zarysu dla zbioru [2], dzięki tabeli 2 oraz wizualizacji tych danych zawartych na wykresie 2 można wywnioskować, że optymalną liczbą klastrów jest liczba z przedziału [5; 8], gdyż dla tego przedziału współczynnik zarysu osiągał najwyższe wartości. Warto zauważyć, że dla liczby klastrów równej 29, dla tego zbioru danych, widać delikatny wzrost dokładności, dlatego zasadne może okazać się wykorzystanie go przy odpowiedniej manipulacji innymi parametrami. Natomiast badanie wpływu wartości maksymalnej liczby iteracji wskazało, że największy wpływ na wartość współczynnika zarysu ma liczba z przedziału [1e6; 1e11] dla liczby klastrów równej 19 lub 25. Warty zauważenia jet fakt, że dal liczby klastrów równej 8, między wartościami 1e10 a 1e11 zaobserwowano wzrost wartości współczynnika zarysu, co może świadczyć o tym, że dobranie maksymalnej liczby iteracji o podobnej wartości do wymienionych może skutkować wzrostem wartości współczynnika. Niestety dla pozostałych wartości liczby klastrów nie zauważono żadnej wartości maksymalnej liczby iteracji, którą należałoby dokładniej zbadać - nie udało się ustalić rzędu wielkości tego parametru, który mógłby mieć wpływ na końcowy wynik.

Analizując wartości współczynnika zarysu dla zbioru [3], dzięki tabeli 3 oraz wizualizacji tych danych zawartych na wykresie 3 można wywnioskować, że optymalną liczbą klastrów jest liczba naturalna z przedziału [6; 10], gdyż dla tego przedziału współczynnik zarysu osiągał najwyższe wartości. Warto zauważyć, że dla liczby klastrów większej niż 25, dla tego zbioru danych, widać delikatny wzrost dokładności, dlatego zasadne może okazać się wykorzystanie tej wartości przy odpowiedniej manipulacji innymi parametrami. Natomiast badanie wpływu wartości maksymalnej liczby iteracji wskazało, że największy wpływ na wartość współczynnika zarysu ma liczba z przedziału [1e7; 1e101] dla liczby klastrów równej 25. Niestety dla pozostałych wartości liczby klastrów nie zauważono żadnej wartości maksymalnej liczby iteracji, którą należałoby dokładniej zbadać - nie udało się ustalić rzędu wielkości tego parametru, który mógłby mieć wpływ na końcowy wynik.

3.2. Algorytm aglomeracyjny

Rysunki 4, 5 oraz 6 przedstawiają wyniki klasteryzacji trzech zbiorów danych, z wykorzystaniem algorytmu aglomeracyjnego. Podobnie jak w przypadku pozostałych metod grupowania, jako miarę jakości podziału danych przyjęto *współczynnik zarysu*. Każdy z wymienionych rysunków zawiera cztery wykresy, dla czterech różnych metod łączenia (ang. *linkage*). Zmienną zależną na wykresach jest wartość współczynnika zarysu a niezależną - liczba klastrów, na które został podzielony zbiór danych. Dodatkowo trzy z czterech

wykresów zawierają po dwie krzywe, każda związana z inną metryką wykorzystaną przez algorytm aglomeracyjny. Wyjątkiem jest tutaj metoda łączenia Warda, ponieważ wymaga ona wykorzystania metryki euklidesowej, do poprawnego działania. W legendzie wykresów zamieszczona została również maksymalna (czyli najlepsza) wartość współczynnika zarysu, jaką udało się dla danych parametrów uzyskać. Tak więc w sumie zbadana została jakość grupowania danych przez algorytm aglomeracyjny w zależności od trzech różnych parametrów:
— liczba klastrów
— metoda łączenia
— metryka

Pierwszym zbiorem danych poddanym klasteryzacji jest zbiór „Iris Species" . Z rysunku 4 można łatwo wywnioskować, że najlepsze wyniki daje podział na 2 klastry, co dla wszystkich metod łączenia z wyjątkiem *complete* pozwala osiągnąć wartość współczynnika zarysu równą 0.69. Jest to wartość „dobra", oznacza, że udało się zdecydowania podzielić zbiór na odseparowane od siebie grupy, nie oznacza to jednak, że jest to podział najlepszy. Wiemy zresztą z góry, że różnych gatunków irysów w tym zbiorze są trzy. Mogły wydarzyć się różne rzeczy, przykładowo dwie klasy irysów zostały sklejona, lub jedna rozbita na części i przyporządkowana do dwóch pozostałych. Warto mieć na uwadze, że zadanie klasteryzacji jest o tyle specyficzne, że trudno tak naprawdę ocenić jakość podziału danych na grupy, zwłaszcza wtedy, kiedy nie wiadomo ile tych grup powinno być. W przypadku wyników związanych z tym zbiorem można, między innymi, zwrócić jeszcze uwagę na dwie sprawy. Po pierwsze pojedyncza metoda łączenia, dla większej liczby klastrów, osiąga znacznie gorsze wyniki niż pozostałem metody - współczynnik zarysu schodzi nawet poniżej zera. To pokazuje jej wrażliwość na dane odstające i tendencję do tworzenia mniej sferycznych struktur. Po drugie ciekawe są wyniki na drugim wykresie w pierwszym rzędzie, gdziem można dostrzec, że dla metryki miejskiej, najlepszą wartość współczynnika zarysu daje podział na 3 klastry, co jest zgodne z faktycznym podziałem tego zbioru. Można by wysnuć jeszcze wiele ciekawszych wniosków, gdyby zastosować dla tego zbioru miary zewnętrzne, pozwalające uwzględnić wiedzę o rzeczywistych klasach próbek.

Na rys. 5 dzieje się już znacznie więcej, niż w przypadku zbioru danych irysów. Widać tutaj wyraźnie, że podział na dwie grupy nie jest najlepszym pomysłem. Nie możemy natomiast skonfrontować tego spostrzeżenia z faktycznym podziałem danych, gdyż nie jest on znany. Stajemy więc przed najbardziej rzeczywistym problemem klasteryzacji, rozwiązywanym w ramach tego zadania. Dla metod łączenia „complete" i „ average" widać, że najlepiej wypada podział danych na 5 grup. W przypadku obu tych metod taki podział prowadzi również do tej samej wartości współczynnika zarysu, jaką jest 0.44. Pomijając fakt, że jest to najlepsza osiągnięta wartość w tym zadaniu, to w ogóle nie jest to wynik, który można uznać za bardzo dobry. Należy jednak pamiętać, że wartość tej miary może być zupełnie różna nawet dla dobrych podziałów zbiorów, ale o różnej charakterystyce. Warto jeszcze w tym momencie podkreślić, że jest dość duże podobieństwo między metodami łączenia „complete" i „average", co widać w podobieństwie wykresów dru-

giego i trzeciego. Trochę inaczej zachowuje się metoda łączenia Warda, która pozwala osiągnąć najwyższą wartość współczynnika zarysu dla 6 klastrów, jest ona jednak wciąż bardzo zbliżona do wspomnianej wcześniej, najlepszej otrzymanej wartości. W tym przypadku metoda łączenia „single" nie sprawuje się zbyt dobrze. Jak widać nie odpowiada charakterystyce zbioru, gdzie próbki są być może pogrupowane w bardziej sferyczne struktury. Na końcu można jeszcze wspomnieć, że obie metryki sprawują się podobnie, zwłaszcza w przypadku optymalnej liczby klastrów.

Ostatni zbiór danych jest, w przeciwieństwie do dwóch poprzednich, jedynie dwuwymiarowy. Jest też zupełnie sztucznie wygenerowany. Natomiast jego kształt, który przypomina dwa nachodzące na siebie półksiężyce, pozwala wyciągnąć wiele ciekawych wniosków na temat metody aglomeracyjnej i samego współczynnika zarysu. Po pierwsze należy stwierdzić, że zgodnie z wykresami najgorzej sprawuje się metoda łączenia „single". Nie jest to jednak zgodne z prawdą, gdyż metoda ta jako jedyna, pozwala faktycznie podzielić zbiór danych na dwa wspomniane półksiężyce. Grupy te nie mają sferycznego kształtu, więc metoda ta nadaje się do nich najlepiej. Niestety współczynnik zarysu nie potraktuje takiego podziału jako najlepszy, gdyż jest oparty na średnich odległościach między próbkami. Pozostałem metody łączenia dają wyniki podobne, to znaczy współczynnik zarysu na poziomie 0.5, czyli „rozsądny", a optymalna liczba klastrów to 7 lub 8. Wynik taki oznacza stworzenia w ramach obu półksiężyców niewielkich, sferycznych grup. Dla takiego podziału współczynnik zarysu jest znacznie większy niż dla podziału na rzeczywiste klasy.

3.3. Algorytm EM

Podczas ewaluacji wyników przeprowadzonych eksperymentów pierwszym parametrem, który został poddany analizie była liczba klastrów. Na początku program został uruchomiony przy stałej wartości maksymalnej liczby iteracji (300) dla liczby klastrów przyjmującej wartość od 1 do 20. Na podstawie wyników z wstępnego eksperymentu zostało wybranych 5 wartości (4,6,9,15,20) przy których zauważono największe "załamania" wykresu i które posłużyły jako wartości parametru dla kolejnych testów.

Dla każdego zbioru optymalna liczba skupisk przyjmuje inną wartość, co jest związane z samą charakterystyką danych użytych do eksperymentu. Dla zbioru Irysów najlepszą wartością klastrów wydaje się być 4, w przypadku Customers jest to 6, a dla zbioru danych Moons najlepsze wyniki były osiągane przy 9 klastrach. Dla każdego zbioru danych zwiększenie liczby klastrów ponad wartość optymalną powodowało coraz częstsze występowanie ujemnych wartości współczynnika zarysu wśród testowanych konfiguracji maksymalnych iteracji oraz macierzy kowariancji. Różnice między wynikami w ramach sąsiednich maksymalnych liczb iteracji również wydawały się większe.

Kolejnym parametrem, który ulegał zmianie podczas przeprowadzonych testów algorytmu była liczba maksymalnych iteracji. Maksymalna liczba iteracji została dobrana poniekąd wykładniczo, aby możliwa była analiza zachowania metody w przypadku zarówno małych, jak i skrajnie dużych wartości. W przypadku każdego zbioru danych można zauważyć pewną prawidłowość.

Algorytm osiągał zadowalające wyniki nawet w przypadku mniejszych wartości maksymalnych iteracji (100 - 200) i nie ulegały one poprawie dla wartości większych.

Ostatnim z parametrów wykorzystywanym podczas rozwiązania problem setu był rodzaj macierzy kowariancji stosowanej przy metodzie EM. Dostępne były 4 warianty, z których każdy wpływa na zachowanie algorytmu:
— *full*
— *tied*
— *diagonal*
— *spherical*
Przypadek dla każdego zestawu danych należy rozpatrywać indywidualnie, gdyż dla każdego zbioru optymalną wartość osiągał inny rodzaj wariantu macierzy kowariancji.

3.4. Algorytm DBSCAN

W przypadku metryki Euklidesowej dla zbioru [1] można zauważyć tendencję, że wraz ze wzrostem minimalnej liczby próbek (ang. min samples) wartość współczynnika zarysu (ang. Silhouette score) wzrasta dla tych samych wartości epsilon z początku zakresu tzn od *eps=0.3* do *eps=0.6*. Jest to szczególnie widoczne gdy porównamy tabele 34 oraz 41. Ciekawą obserwacją jest też to, że od wartości *eps=1.0* dla praktycznie wszystkich zbadanych wartości *min samples* wartość *Silhouette score* to *0.6864*. Co więcej dziwną sprawą jest, że dla praktycznie wszystkich wartości *min samples*, zazwyczaj dla drugiej wartości *eps* jest dosyć duży spadek wartości współczynnika zarysu, natomiast dla kolejnych wartości można zaobserwować wzrost wartości współczynnika aż do osiągnięcia wcześniej wspomnianej wartości maksymalnej. Co do samych wartości *eps* to to pomimo zastosowanie liczb z przedziału od *0.1* do *9.9* z częstotliwością *0.1* udało się uzyskać wyniki tylko do wartości *eps=1.6* bez względu na wartość *min samples*. Gdy została użyta metryka Manhattan to do osiągnięcia maksymalnej wartości współczynnika zarysu wymaga była wyższa wartość *eps* natomiast było możliwe wyznaczenie klastrów dla większych wartości *eps* tzn. do wartości *2.6* zamiast *1.6*.

W przypadku zbioru [2] dla metryki Euklidesowej wartość współczynnika zarysu wzrastała wraz z wartością *eps*. Warto tutaj wspomnieć o zmianie w wartościach *eps*, które zostały opisane w sekcji 1. W większości przypadków był to wzrost w miarę liniowy. Najlepszy wynik udało się uzyskać dla *min samples=7*. Gdy została użyta metryka Manhattan to od razu można zauważyć to, że przyrost wartości *Silhouette score* jest mniej liniowy i bardziej *'poszarpany i kwadratowy'* co można tłumaczyć charakterem działania metryki. Same wyniki są dosyć porównywalne natomiast jest tutaj zachowana ta tendencja co dla zbioru [1], że wymagane są większe wartości *eps* do uzyskania podobnych wartości współczynnika.

Ostatnim użytym zbiorem jest [3], dla metryki Euklidesowej bardzo ciekawą sprawą jest to, że bez względu na wartość *min sample*, wartość współczynnika zarysu udało się wyliczyć tylko dla wartości *eps* równej *0.1* oraz *0.2* oraz dla *eps=0.2* wynik jest zawsze taki sam. Jedyne różnice można zauważyć dla *eps=0.1* wraz ze wzrostem *min samples* wzrasta wartość *Silhouette score*.

Jest to wybitnie widoczne *min samples=9* gdy wyjątkowo jest wykres funkcji liniowej, która maleje. Jest to spowodowane tym, że właśnie tam udało się uzyskać najlepszy grupowanie a wartość współczynnika zarysu była równa *0.3678*. Gdy została użyta metryka Manhattan to wyniki dla *eps=0.1* wyniki diametralnie się pogorszyły, natomiast dla wartości *eps=0.2* uzyskano takie samo maksimum jak dla pierwszej metryki.

4. Wnioski

Podsumowując wykonane zadanie wnioskujemy, że:

— W algorytmie aglomeracyjnym nie ma tak dużego znaczenia wybór metryki, natomiast metoda łączenia jest decydująca dla poprawnego działania algorytmu

— Metoda łączenia „single" algorytmu aglomeracyjnego zazwyczaj sprawuje się gorzej, jest bardziej podatna na kształt danych i dane odstające, ze zwzględu na swój charakter prowadzi zazwyczaj do niższych wartości współczynnika zarysu, pozostałe metody łączenia sprawują się podobnie i prowadzą do podobnych podziałów o sferycznym kształcie

— Sam współczynnik zarysu nie wystarczy aby dobrze ocenić jakość klasteryzacji, jest on bardzo zależny od charakterystyki zbioru danych

— Dla metody *k-średnich* najważniejszym parametrem jest liczba klastrów, która ma największy wpływ na wartość współczynnika zarysu. Maksymalna liczba iteracji ma drugorzędne znaczenie, jednakże może ona wpłynąć korzystnie na wartość współczynnika zarysu, zwykle jest to zaobserwowane dla dużych wartości liczby iteracji, efekt ten jest lepiej widoczny dla badania, w którym wykorzystana została większa liczba klastrów.

— Dla metody *EM* wybór liczby klastrów ma kluczowe znaczenie dla uzyskiwanych wartości współczynnika zarysu. Maksymalna liczba iteracji ma natomiast znikomy wpływ na otrzymane wyniki, jednak tak jak w przypadku algorytmu k-średnich może one wpłynąć pozytywnie na wartość współczynnika zarysu

— Dla metody *EM* optymalne konfiguracje różnią się między wybranymi zbiorami danych, dlatego ważne jest dobranie ich indywidualnie dla każdego zestawu danych

— Dla metody *DBSCAN* wybór metryki nie wpływa znacząco na wyniki natomiast dla metryki Manhattan po prostu potrzebne są większe wartości parametru *eps*

— Dla metody *DBSCAN* i wybranych zbiorów warto stosować większe wartości parametru *min samples*. Samo znajdowanie wartości warto przeprowadzać dla danego zbioru gdyż pominiecie tego kroku może pogorszyć wyniki

Literatura

[1] https://www.kaggle.com/uciml/iris

[2] https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

[3] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html