Sprawozdanie PD4 WdUM

Jakub Niemyjski

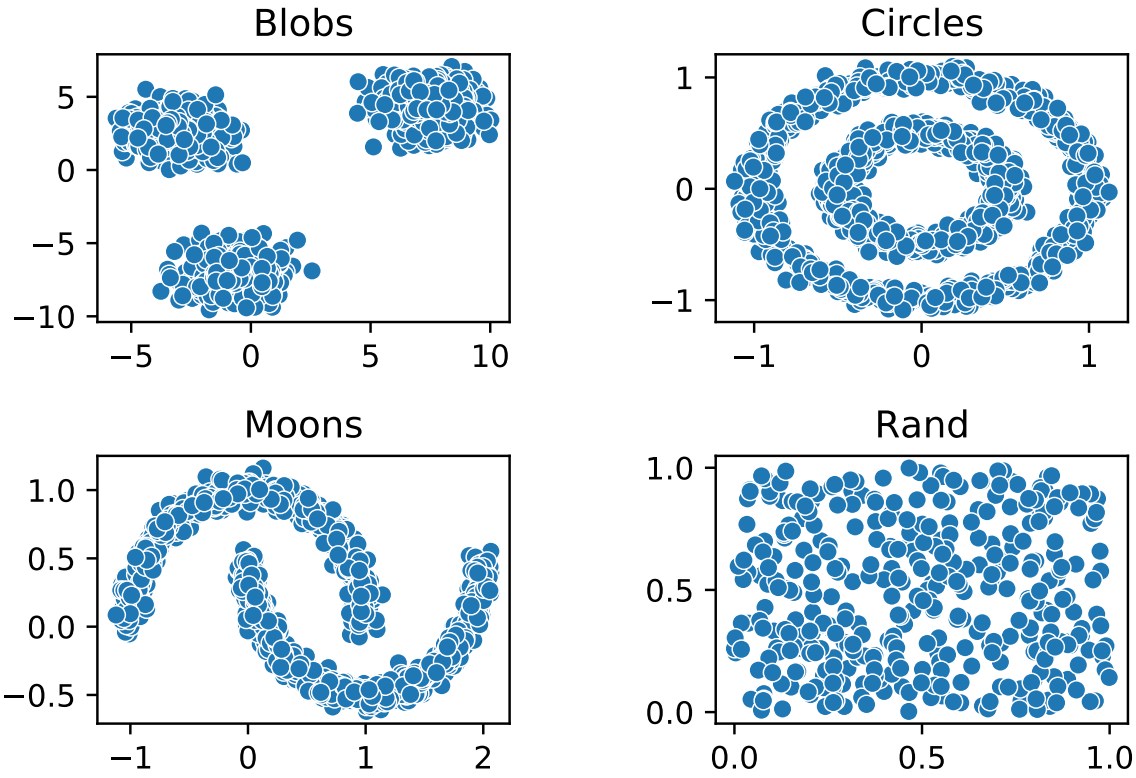
27 maja 2024

1 Cel

Praca ta ma na celu przybliżenie metod analizy skupień: *k-średnich, metody hierarchicznej* oraz *DBSCAN*. Bazując na załączonych danych chcemy odpowiedzieć na postawione pytania badawcze.

2 Dane

Rozważanymi danymi są wygenerowane sztucznie za pomocą biblioteki *sklearn* 3 typy zbiorów złożonych z 1500 obserwacji każdy, utworzone kolejno przez funkcje `make_blobs , make_circles , make_moons` z pakietu `sklearn.datasets`. Ostatni, czwarty zbiór jest wygenerowany korzystając z dwuwymiarowego rozkładu jednostajnego na przedziale $[0, 1]$. Ilustracje tychże zbiorów umieszczone są poniżej. Niestety, prawdopodobnie ze względu na to, że na rysunku w prawym dolnym rogu jest zbyt dużo obserwacji, format pdf nie wyświetla ich poprawnie, pomimo, że kod w Pythonie wygenerował poprawny rysunek, co widać w kodach źródłowych do raportu.



Rysunek 1: Prezentacja rozważanych zbiorów.

3 Problemy badawcze

1. Które algorytmy wymagają liczby klastrów na wejściu? Jak wybór tego parametru wpływa na wyniki?

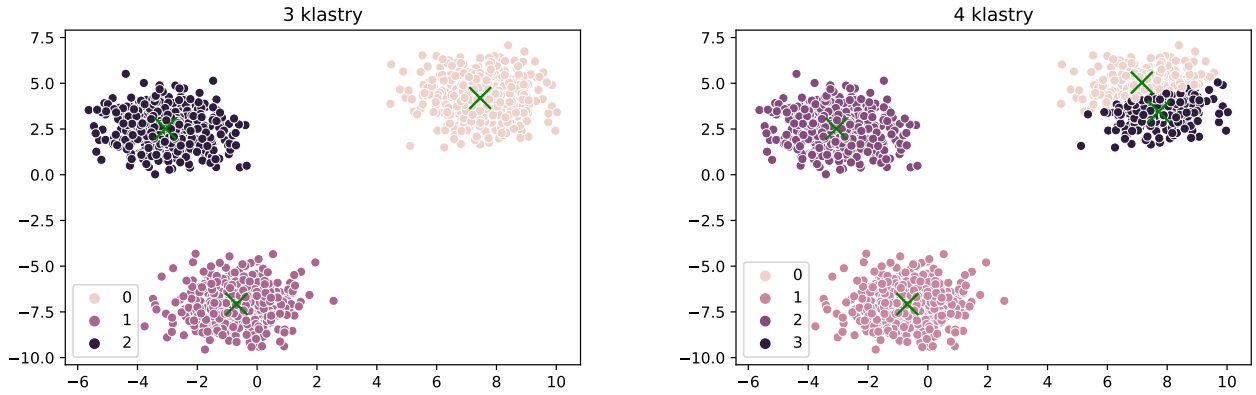
2. Zbadaj wpływ parametru szumu (noise) dla zestawów danych 2 i 3.

3. Jaka jest zależność między całkowitą sumą odległości między punktami w klastrach, a liczbą klastrów?

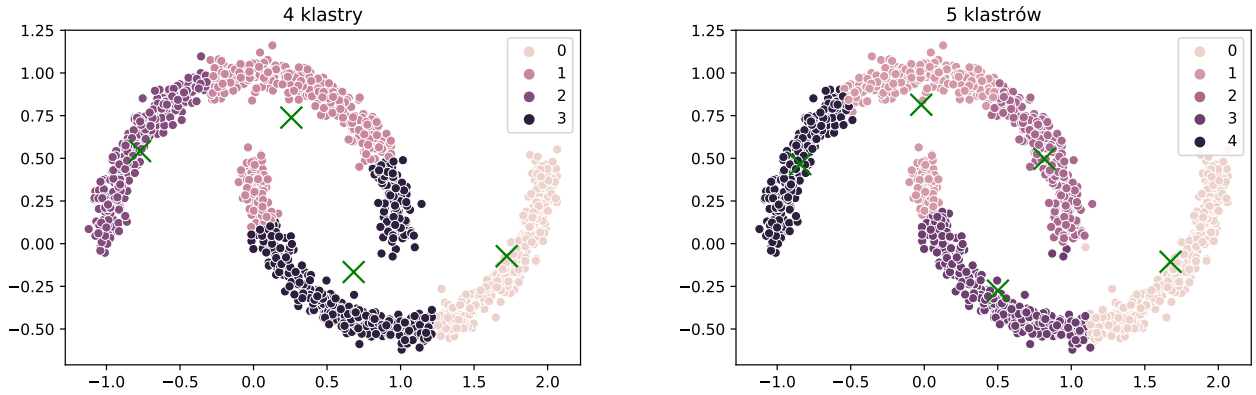
4 Problem 1

4.1 KMeans

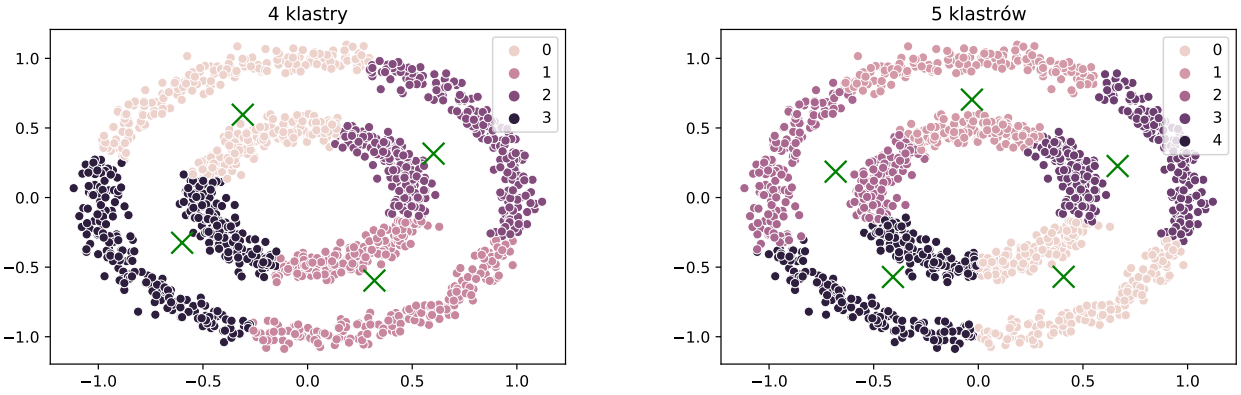
Algorytm k-średnich na wejściu wymaga zdecydowania się na ustaloną liczbę klastrów. Na rysunkach (2)-(5) można zaobserwować, jak różna ich liczba może wpływać na przyporządkowanie obserwacji do klastrów.



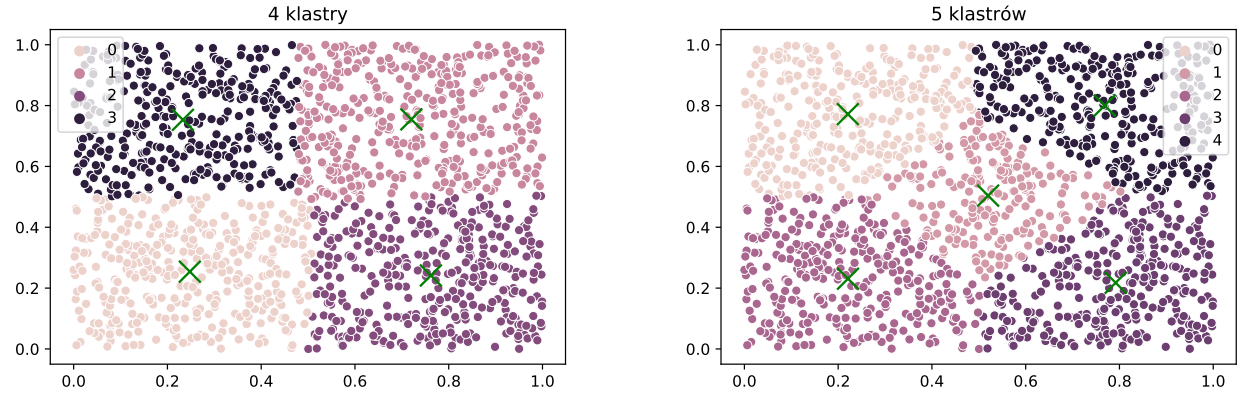
Rysunek 2: Blobs w metodzie KMeans



Rysunek 3: Księżyce w metodzie KMeans



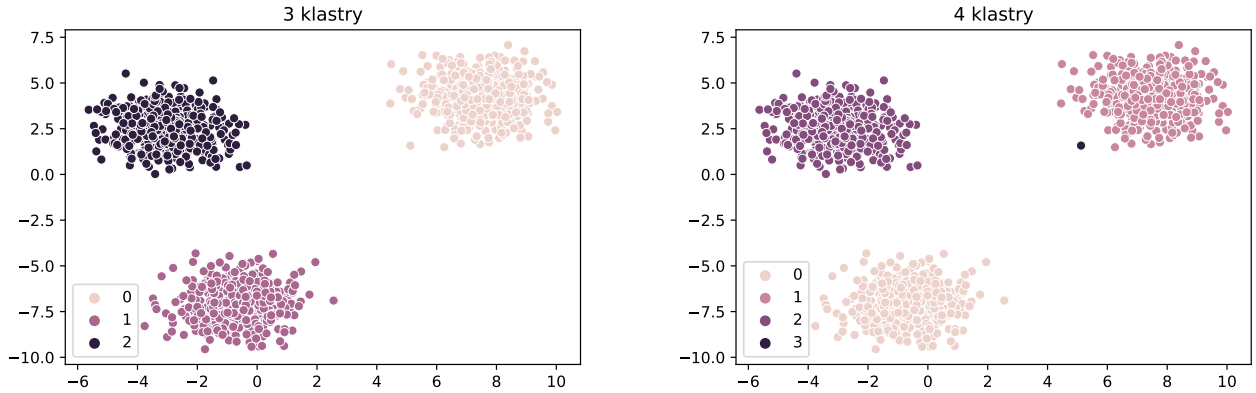
Rysunek 4: Elipsy w metodzie KMeans



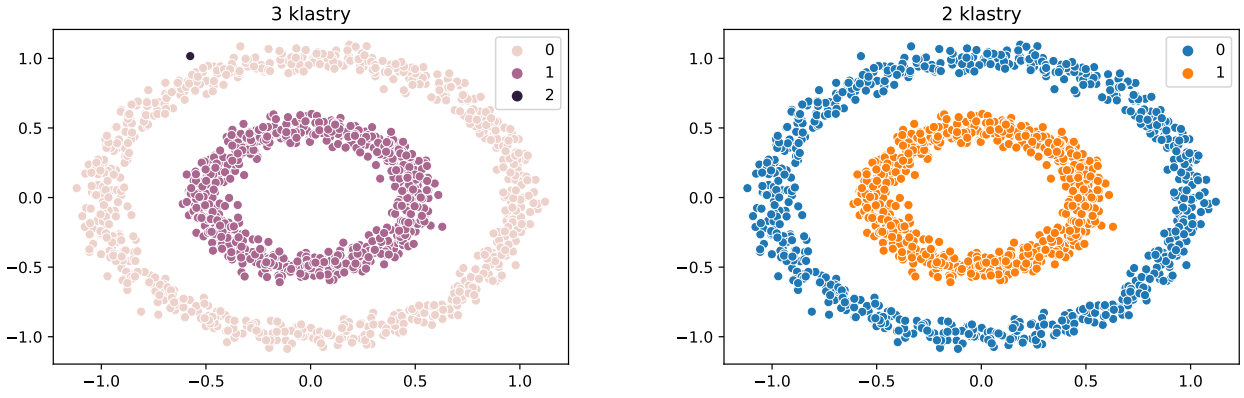
Rysunek 5: Jednostajny rozkład punktów w metodzie KMeans

4.2 Metoda hierarchiczna

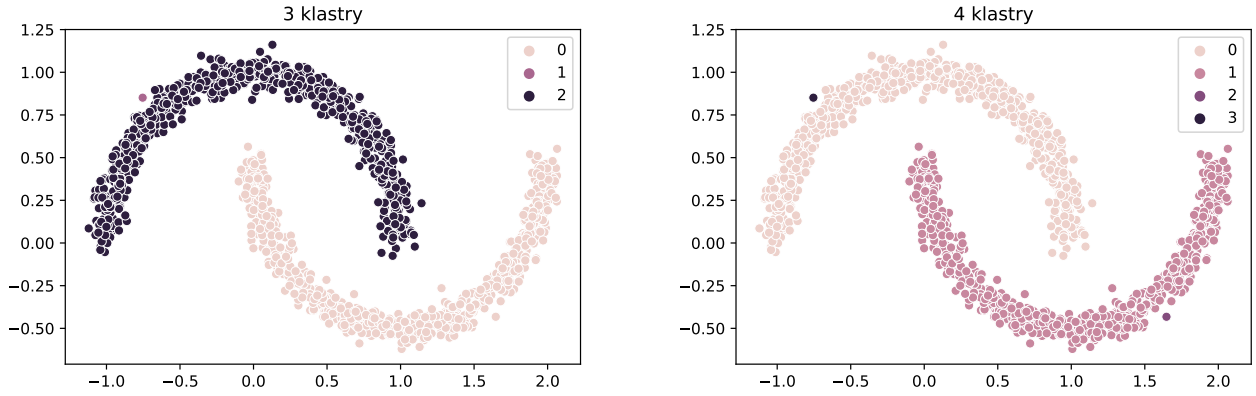
Metody hierarchiczne tworzą podział zbioru od największej liczby podziałów (czyli, gdy każda obserwacja jest sama w sobie samotną grupą) aż do jednej grupy złożonej ze wszystkich obserwacji. W wyborze optymalnej liczby grup, na której powinno się zaprzestać dalsze łączenie, pomagają dendrogramy, o których mowa w sekcji 6. Zatem metody hierarchiczne wymagają na wejściu ustalenia liczby klastrów.Na rysunkach (6)-(9) widać jak metody hierarchiczne łączą w grupy nasze zbiory.



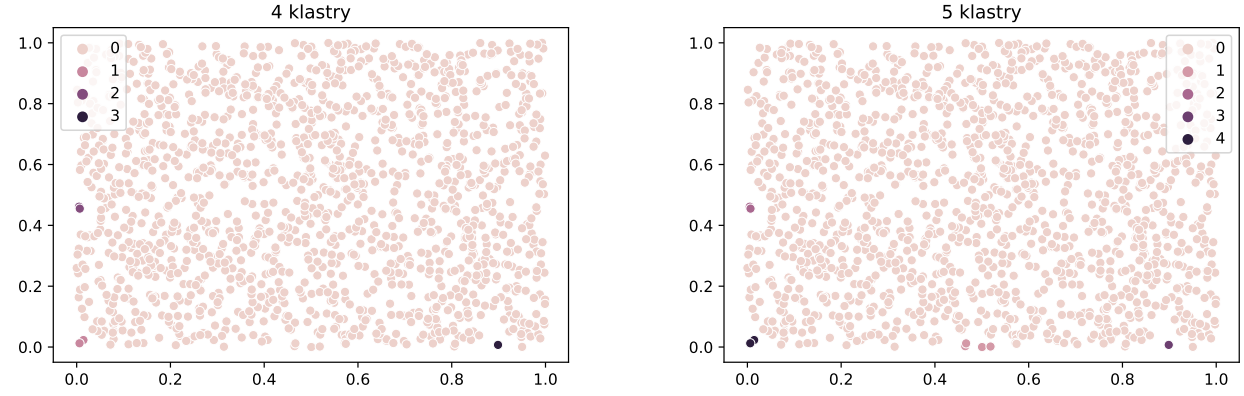
Rysunek 6: Blobs w metodzie hierarchicznej.



Rysunek 7: Circles w metodzie hierarchicznej.



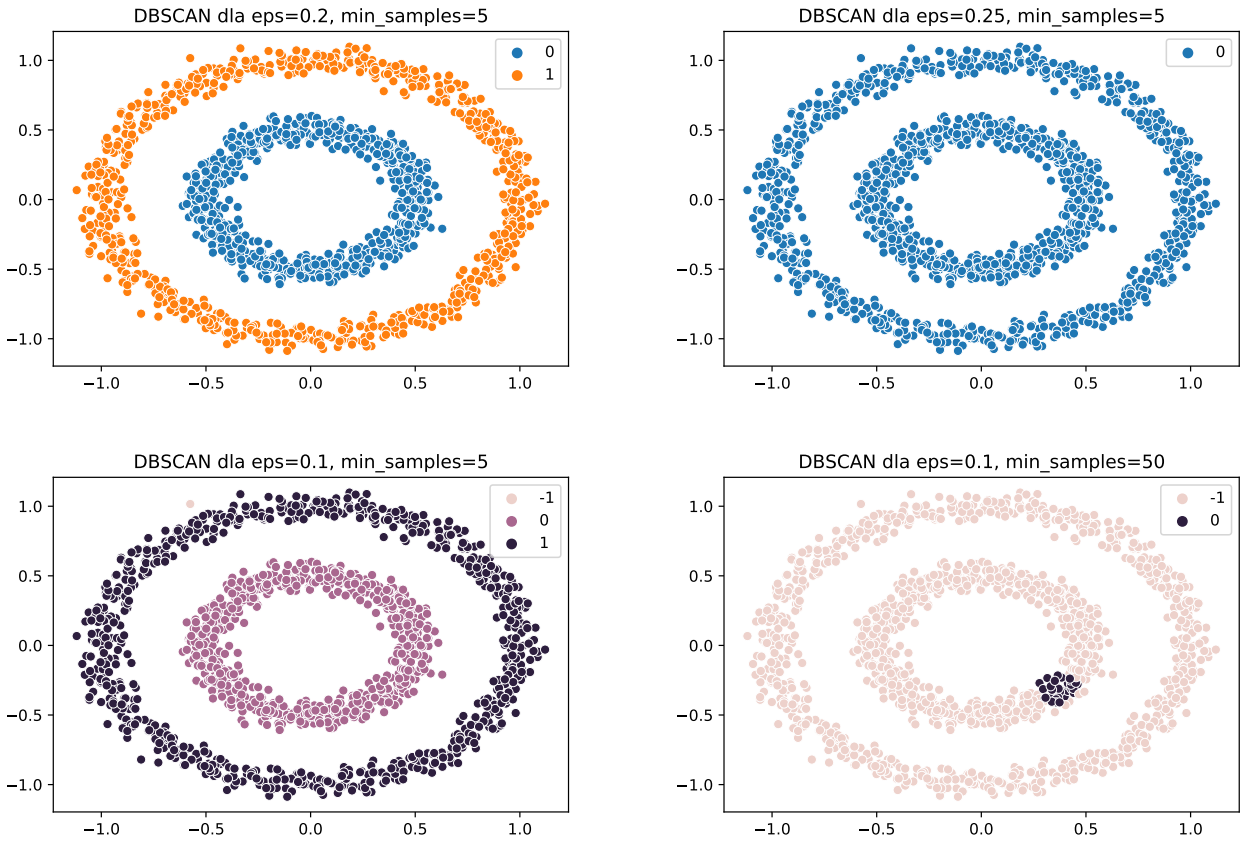
Rysunek 8: Moons w metodzie hierarchicznej.



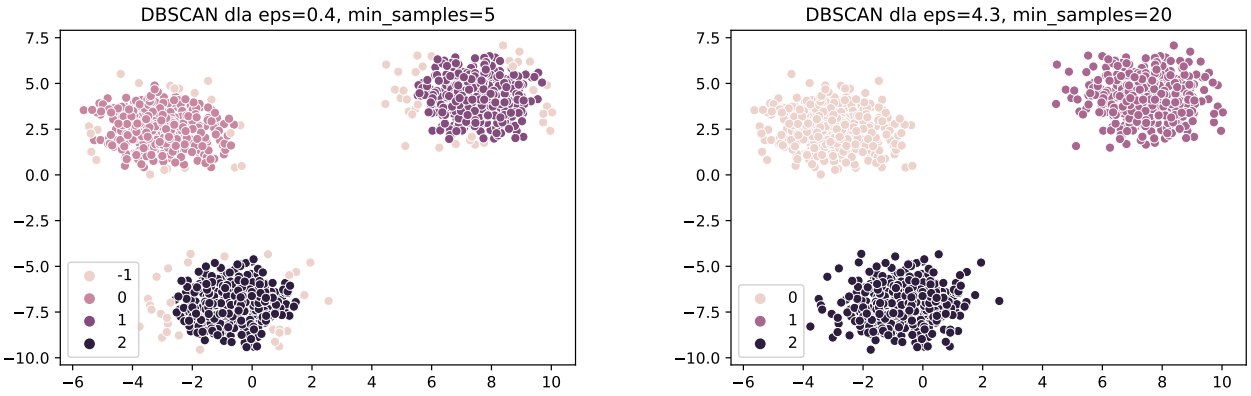
Rysunek 9: Zbiór losowany z rozkładu jednostajnego w metodzie hierarchicznej.

4.3 DBSCAN

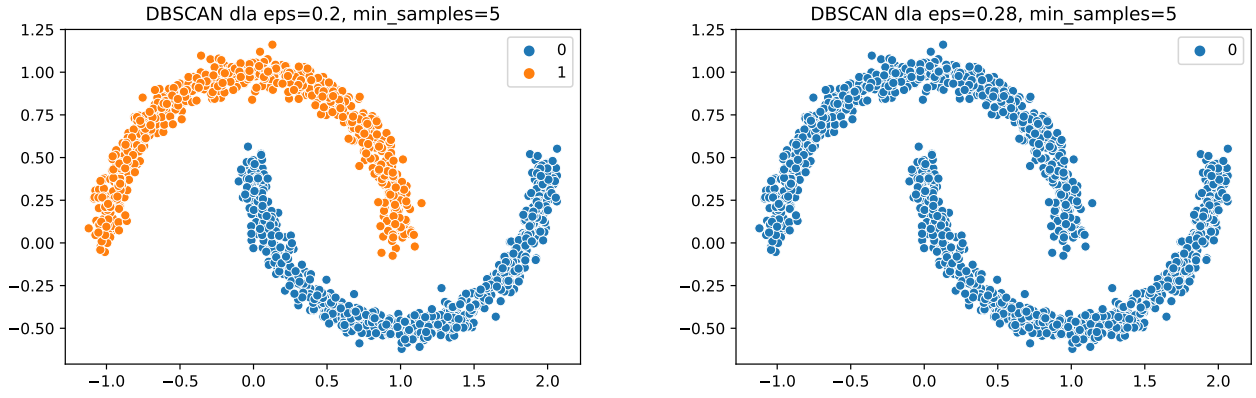
Jest to metoda klasteryzacji, która nie przyjmuje na wejściu liczby klastrów, ale między innymi *eps* oraz *min_samples*, które odpowiadają za promień kuli, w której to szukamy kolejnych punktów, które przypiszemy do tej samej grupy punktów co tzw. *core point*. Jednakże punkty te przypisywane są do tej grupy pod warunkiem, że w tej kuli jest co najmniej *min_samples* obserwacji. Na rysunkach (10)-(13) zostało zobrazowane jej działanie dla różnych zbiorów i różnych parametrów.



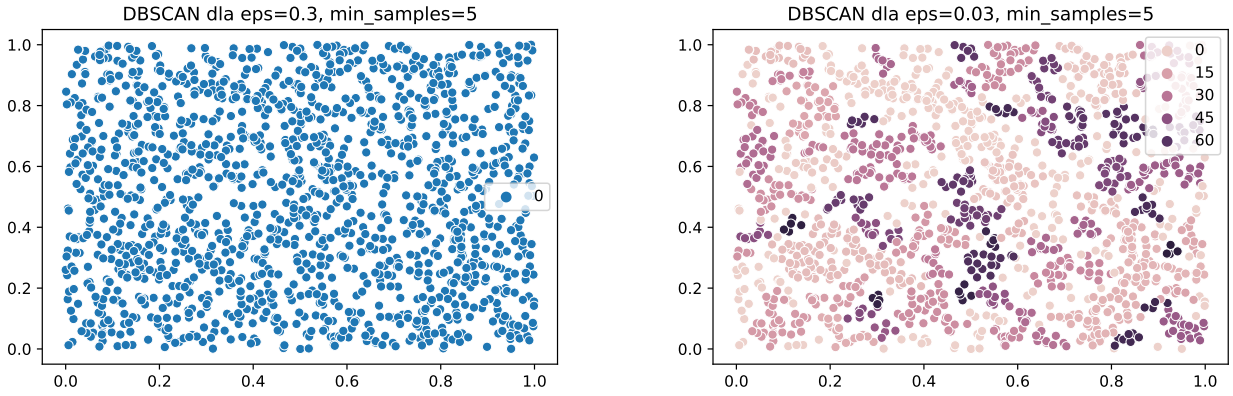
Rysunek 10: Wykresy przedstawiające 1500 obserwacji wygenerowanych metodą `make_circles` i sklasyfikowane metodą DBSCAN z różnymi parametrami.



Rysunek 11: Blobs rozkład punktów w metodzie DBSCAN



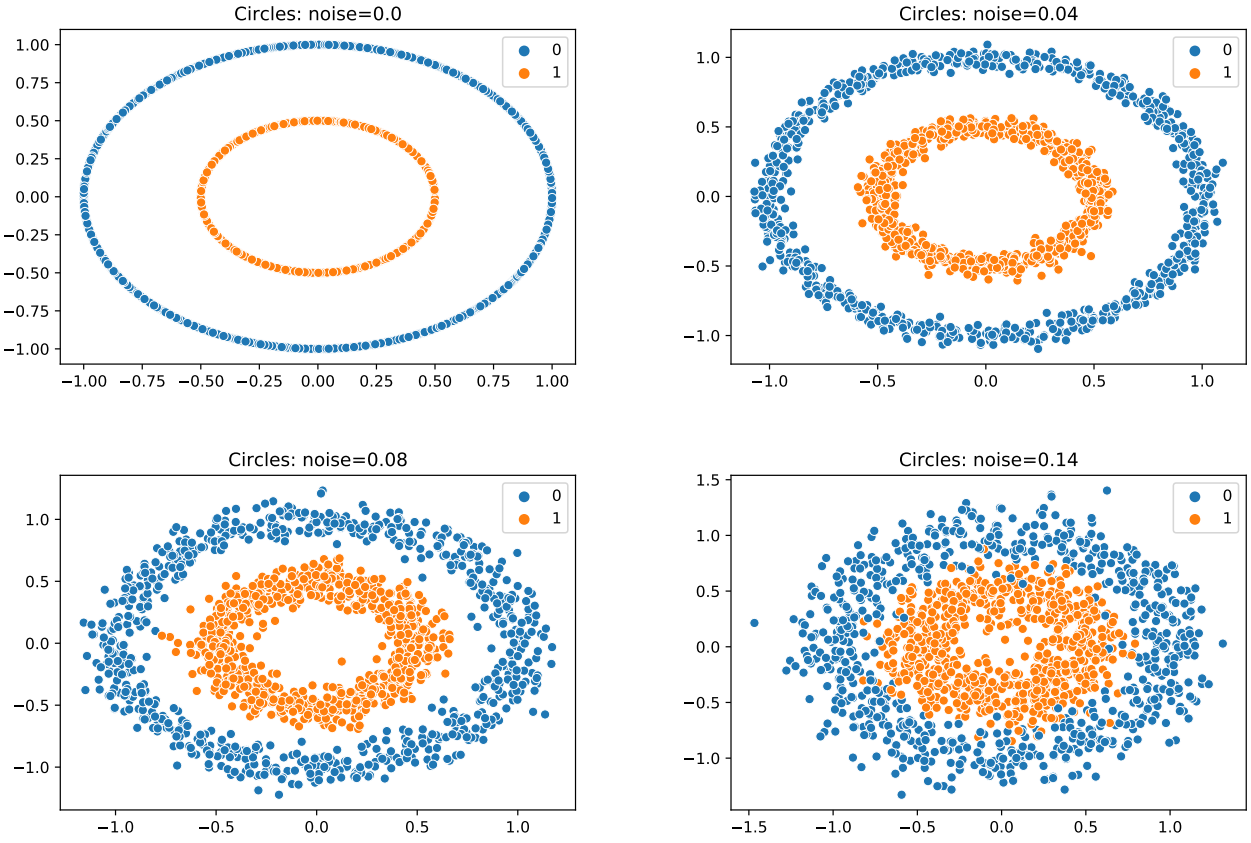
Rysunek 12: Księżyce rozkład punktów w metodzie DBSCAN



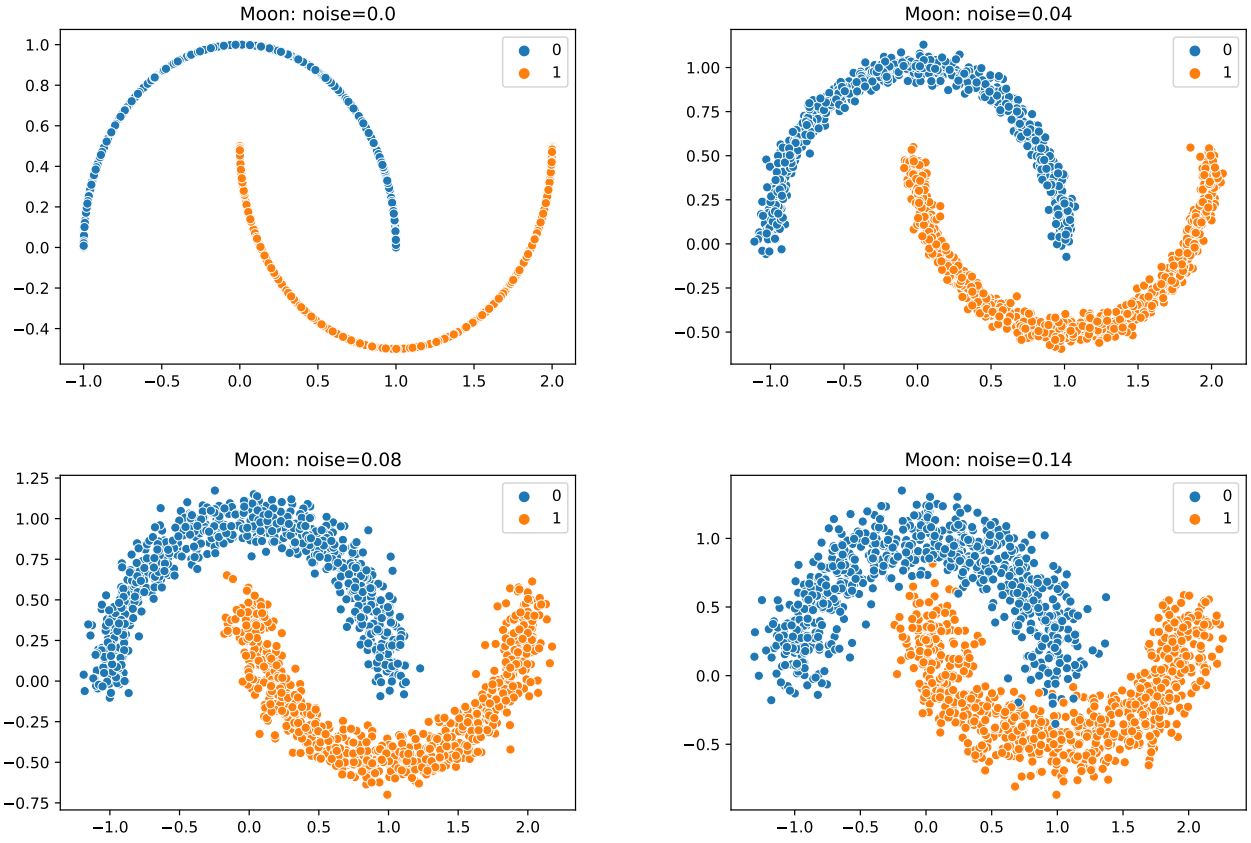
Rysunek 13: Jednostajny rozkład punktów w metodzie DBSCAN

5 Problem 2

Funkcje generujące zbiory drugi i trzeci (czyli `make_circles` i `make_moons`) jako jeden z parametrów przyjmują *noise*. Jest to parametr służący do dodania do wygenerowanych punktów tworzących kółka i księżyce szumu gaussowskiego. Naturalnie, spodziewać się można, że wygenerowane zbiory bez żadnego szumu będą perfekcyjnie formować kształty elips i księżyców, a wraz ze wzrostem szumu, wykresy te będą coraz mniej przypominać te kształty. W celu empirycznej weryfikacji tej tezy, utworzone zostały dwie animacje które sprawdzają jak zmienia się wykres przedstawiony na płaszczyźnie wraz ze wzrostem parametru *noise*. Jedna animacja dla zbioru z elipsami, druga, dla zbioru z księżycami. Kilka momentów tych animacji zostało umieszczonych na wykresach 14 i 15.



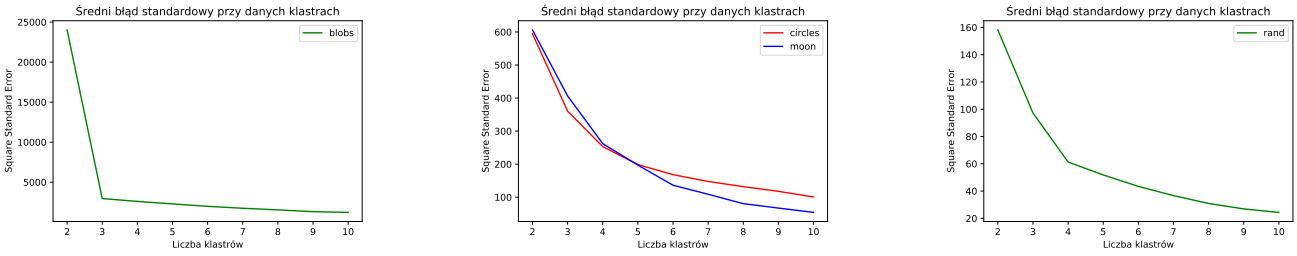
Rysunek 14: Wykresy przedstawiające 1500 obserwacji wygenerowanych metodą `make_circles` dla różnych parametrów *noise*.



Rysunek 15: Wykresy przedstawiające 1500 obserwacji wygenerowanych metodą `make_moons` dla różnych parametrów *noise*.

6 Problem 3

Naturalnie, wraz ze wzrostem liczby klastrów, całkowita suma odległości między punktami w klastrach maleje do zera. Zero jest przyjmowane, jeżeli klastrów jest tyle samo, ile punktów. W metodzie k-średnich, na rysunku 16a widać, że przy przejściu z trzech klastrów do czterech łamana drastycznie zmienia swoje nachylenie, co może wskazywać na to, że liczba trzech klastrów jest tutaj najbardziej optymalna. W reszcie zbiorów, tj. zobrazowanych na rysunkach 16b i 16c mamy już mniej widoczne granice wyboru. Takie obserwacje zgadzają się z tym, co widać na rysunkach 1.

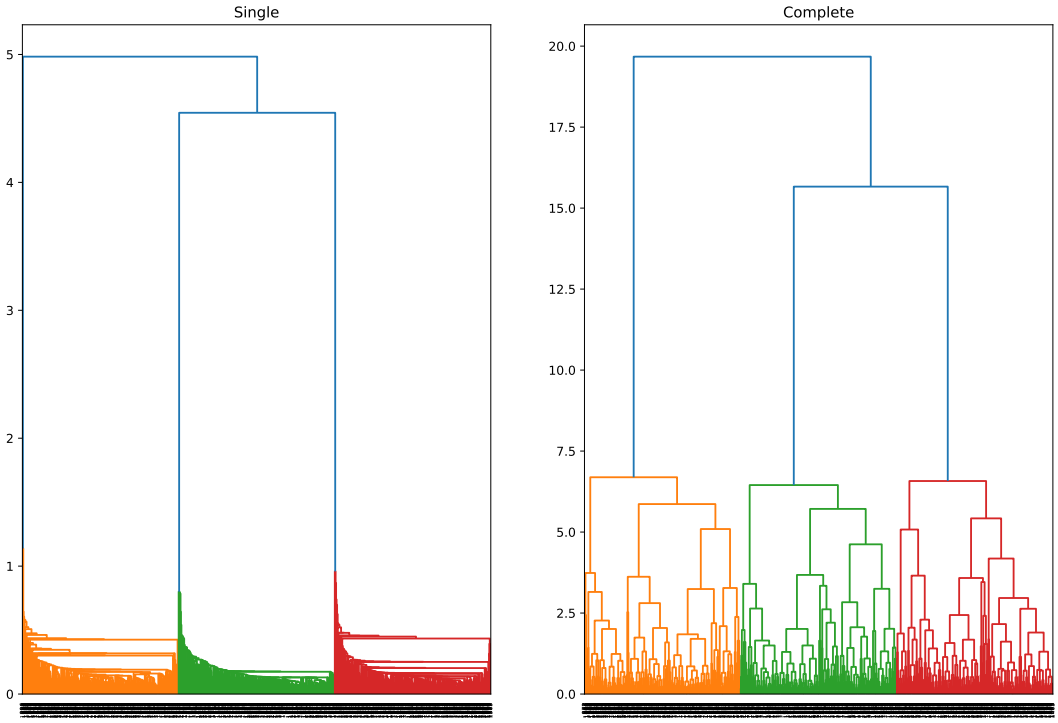


(a) (b) (c)

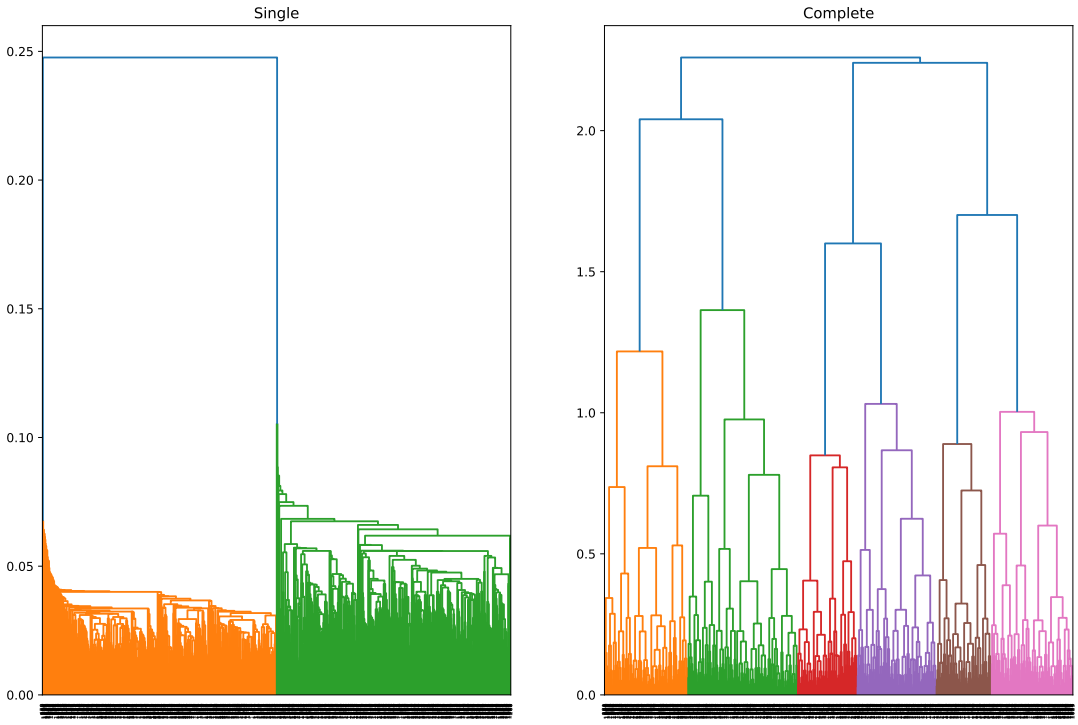
Rysunek 16: SSE dla metody KMeans dla badanych czterech zbiorów.

W celu optymalnego doboru liczby klastrów w metodach hierarchicznych, na rysunkach (17)-(20) sporządzono dendrogramy. Dla zbioru *blobs* widać, że już przy niskiej wartości na osi y na obu wykresach wyklarowane zostały trzy klastry, choć w przypadku łaczenia typu *'single'* zostało to bardziej wyraziście przedstawione.
W zbiorze *moons* i *circles* w metodzie *'single'* ponownie dla małego progu już widać dwa klastry, których liczba jest zgodna z graficzną reprezentacją tych zbiorów na rysunku 1.

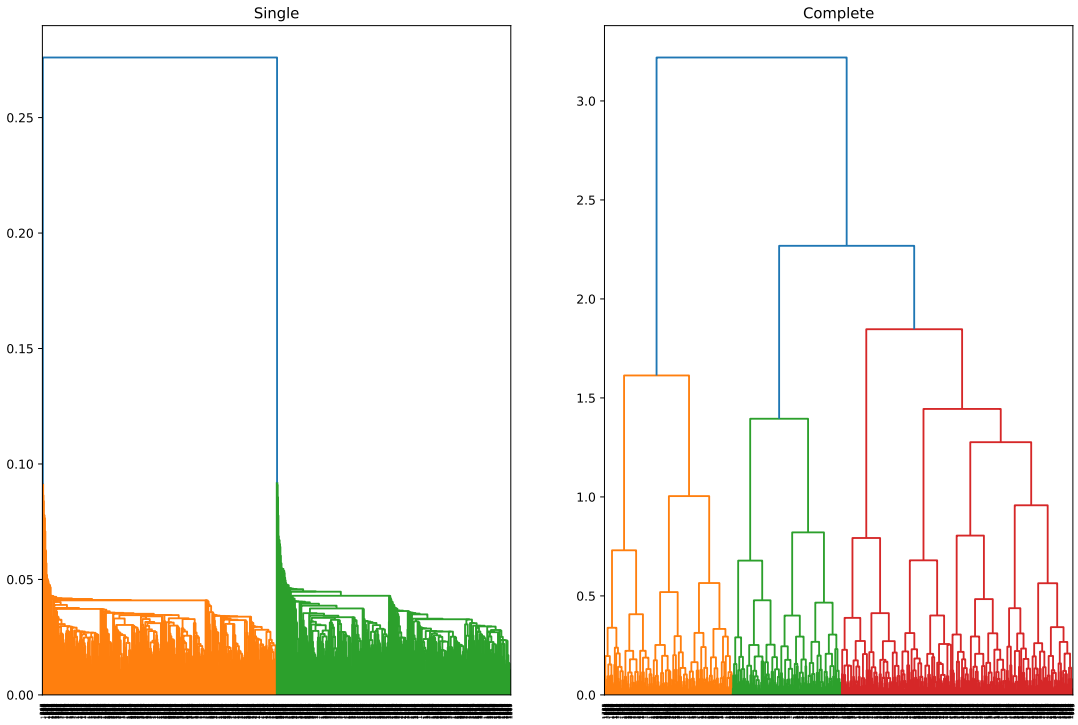
Na rysunku 20 widać, że dla dużej wartości na osi y nadal nie zostało rozstrzygnięte jaka liczba klastrów będzie optymalna, nadal jest ich dużo.



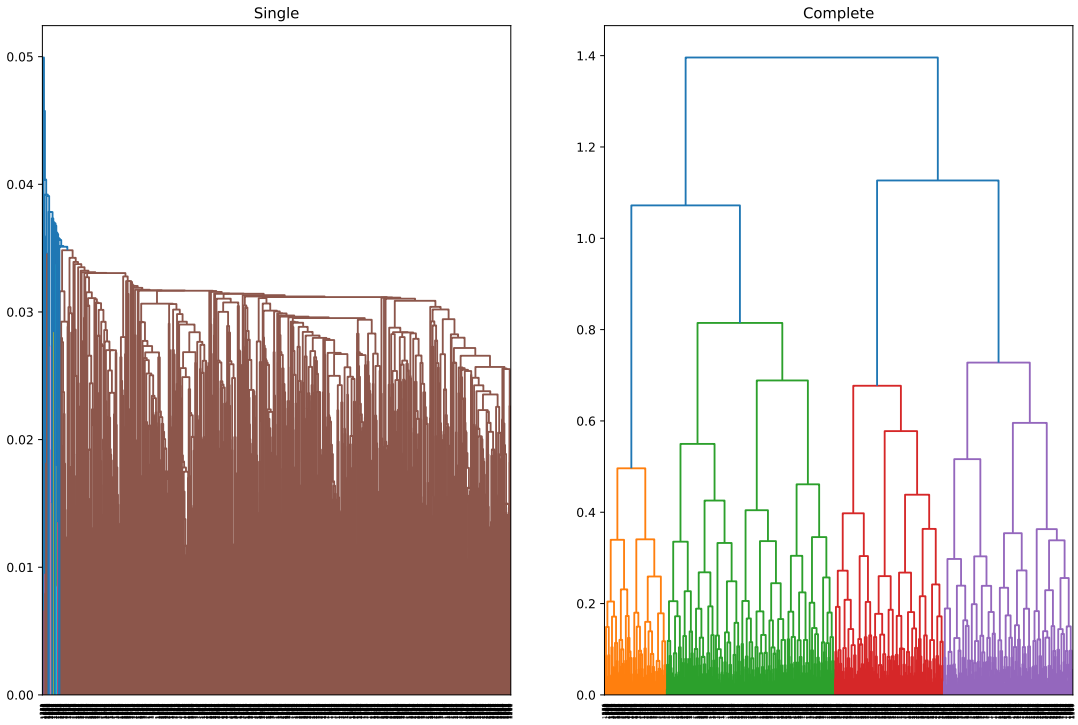
Rysunek 17: Dendrogramy dla zbioru *blobs*.



Rysunek 18: Dendrogramy dla zbioru *circles*.



Rysunek 19: Dendrogramy dla zbioru *moons*.



Rysunek 20: Dendrogramy dla zbioru wylosowanego z rozkładu jednostajnego.

7 Podsumowanie

W tej pracy przybliżone zostały trzy metody nienadzorowanego uczenia maszynowego. Metoda k-średnich, metody hierarchiczne i DBSCAN. Różnią się ona działaniem, bazują na innych metodach porównywania elementów, czy też grup elementów, ale różnią się też zastosowaniem. Jednakże każda z nich może być w pewnych sytuacjach przydatna.