Algorytmy Analizy skupień w praktyce

Jakub Nowacki

1. Opis zbioru danych

Zbiór danych Red Wine Quality zawiera informacje o 1599 gatunkach czerwonego portugalskiego wina. Każdy obiekt w zbiorze danych reprezentuje jedno wino i jest opisany przez 12 zmiennych.

Atrybuty:

- fixed acidity - stała kwasowość, większość kwasów związanych z winem lub stałych lub nielotnych
- volatile acidity - kwasowość lotna, ilość kwasu octowego w winie, która przy zbyt wysokim poziomie może prowadzić do nieprzyjemnego, octowego smaku
- citric acid - kwas cytrynowy, występujący w niewielkich ilościach, kwas cytrynowy może dodawać winom "świeżości" i smaku
- residual sugar - cukier resztkowy, ilość cukru pozostała po zatrzymaniu fermentacji
- chlorides - chlorki, ilość soli w winie
- free sulfur dioxide - wolny dwutlenek siarki, zapobiega rozwojowi drobnoustrojów i utlenianiu wina
- total sulfur dioxide - całkowity dwutlenek siarki, ilość wolnych i związanych form S02
- density - gęstość
- pH - opisuje kwasowość lub zasadowość wina w skali od 0 (bardzo kwaśne) do 14 (bardzo zasadowe); większość win mieści się w przedziale 3-4 w skali pH.
- sulphates - siarczany, dodatek do wina, który może przyczyniać się do zwiększenia poziomu dwutlenku siarki (S02), który działa przeciwbakteryjnie i przeciwzapalnie.

Rozkład wartości

Wartości atrybutów w zbiorze danych Red Wine Quality mają zróżnicowany rozkład. Niektóre atrybuty, takie jak zawartość kwasowości pierwotnej (fixed acidity) lub całkowita zawartość dwutlenku siarki (total sulfur dioxide), mają rozkład normalny. Inne atrybuty, takie jak zawartość cukru pozostałego po procesie fermentacji (residual sugar) lub zawartość alkoholu (alcohol), mają rozkład skośny.

Rzędy wielkości

Wartości atrybutów w zbiorze danych Red Wine Quality mają różne rzędy wielkości. Niektóre atrybuty, takie jak zawartość kwasowości pierwotnej (fixed acidity) lub zawartość alkoholu (alcohol), mają wartości w skali od 0 do 100. Inne atrybuty, takie jak zawartość chlorku sodu (chlorides) lub zawartość siarczanów (sulphates), mają wartości w skali od 0 do 1000.

	Rzędy wielkości	Liczba unikalnych wartości
fixed acidity	100	96
volatile acidity	10	143
citric acid	10	80
residual sugar	100	91
chlorides	10	153
free sulfur dioxide	100	60
total sulfur dioxide	1000	144
density	10	436
pH	10	89
sulphates	10	96
alcohol	100	65
quality	10	6

Statystyki opisowe:

	fixed acidity	volatile acidity	citric acid	residual sugar	\
count	1599.000000	1599.000000	1599.000000	1599.000000	
mean	8.319637	0.527821	0.270976	2.538806	
std	1.741096	0.179060	0.194801	1.409928	
min	4.600000	0.120000	0.000000	0.900000	
25%	7.100000	0.390000	0.090000	1.900000	
50%	7.900000	0.520000	0.260000	2.200000	
75%	9.200000	0.640000	0.420000	2.600000	
max	15.900000	1.580000	1.000000	15.500000	

	chlorides	free sulfur dioxide	total sulfur dioxide	density	\
count	1599.000000	1599.000000	1599.000000	1599.000000	
mean	0.087467	15.874922	46.467792	0.996747	
std	0.047065	10.460157	32.895324	0.001887	
min	0.012000	1.000000	6.000000	0.990070	
25%	0.070000	7.000000	22.000000	0.995600	
50%	0.079000	14.000000	38.000000	0.996750	
75%	0.090000	21.000000	62.000000	0.997835	
max	0.611000	72.000000	289.000000	1.003690	

	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000
mean	3.311113	0.658149	10.422983	5.636023
std	0.154386	0.169507	1.065668	0.807569
min	2.740000	0.330000	8.400000	3.000000
25%	3.210000	0.550000	9.500000	5.000000
50%	3.310000	0.620000	10.200000	6.000000
75%	3.400000	0.730000	11.100000	6.000000
max	4.010000	2.000000	14.900000	8.000000

2. Grupowanie hierarchiczne

Wybrany algorytm grupowania hierarchicznego to algorytm AHC aglomeracyjny.

```
# Definiowanie metod łączenia i miar odległości do przetestowania
linkage_methods = ["single", "complete", "average"]
distance_metrics = ["euclidean", "cosine"]
```

Algorytm został uruchomiony 6 razy dla różnych kombinacji metod łączenia i miar odległości, które zostały wypisane na powyższym rysunku.

	Metody łączenia i miary odległości	Indeks Daviesa-Bouldina
1	Single linkage, euclidean	0.46781651342878555
2	Single linkage, Cosine	0.6503072951178342
3	Complete linkage, Euclidean	1.3820476297764526
4	Complete linkage, Cosine	1.94600407113734
5	Average linkage, Euclidean	0.8869677267400741
6	Average linkage Cosine	1.3212788383283796

Indeks Daviesa-Bouldina (Davies-Bouldin Index) jest miarą jakości klastrów w analizie skupień. Jest obliczany na podstawie stosunku średnich odległości między klastrami do średnich odległości wewnątrz klastrów. Niższa wartość indeksu Daviesa-Bouldina oznacza lepszą jakość klastrów.

1. Single linkage, Euclidean (0.46781651342878555):
Niska wartość indeksu sugeruje, że klastry są dobrze zdefiniowane i odseparowane od siebie.

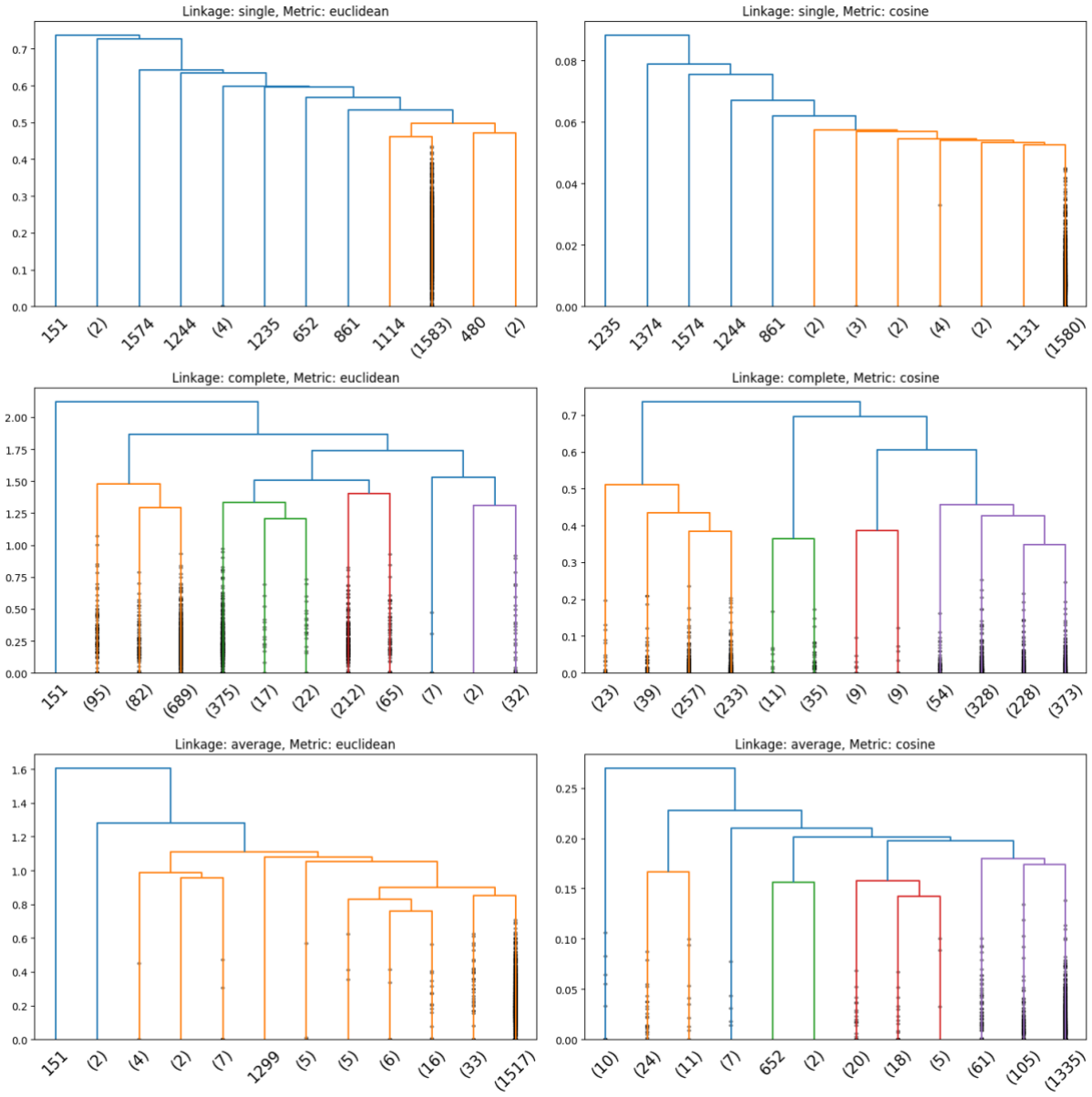
2. Single linkage, Cosine (0.6503072951178342):
Wyższa wartość DBI w porównaniu do Euclidean wskazuje, że dla tej konfiguracji klastry mogą być mniej jednoznaczne lub mniej kompaktowe.

3. Complete linkage, Euclidean (1.3820476297764526):
Wartość DBI jest wyższa, co może oznaczać, że klastry są bardziej rozmyte lub mniej jednorodne.

4. Complete linkage, Cosine (1.94600407113734):
Wartość DBI jest wysoka, co może sugerować, że klastry są mniej jednorodne w przypadku użycia odległości kosinusowej.

5. Average linkage, Euclidean (0.8869677267400741):
Średnia wartość DBI wskazuje na umiarkowaną jakość klastrów z użyciem tej konfiguracji.

6. Average linkage, Cosine (1.3212788383283796):
Wysoka wartość DBI w odniesieniu do Euclidean sugeruje, że dla tej konfiguracji klastry mogą być bardziej rozproszone



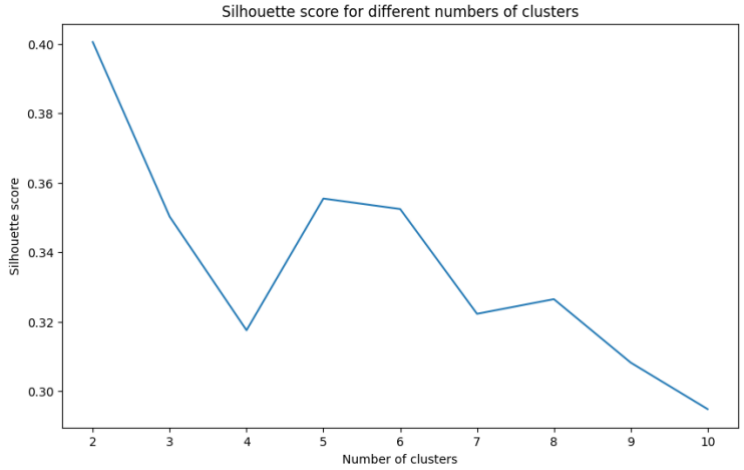
3. Grupowanie podziałowe

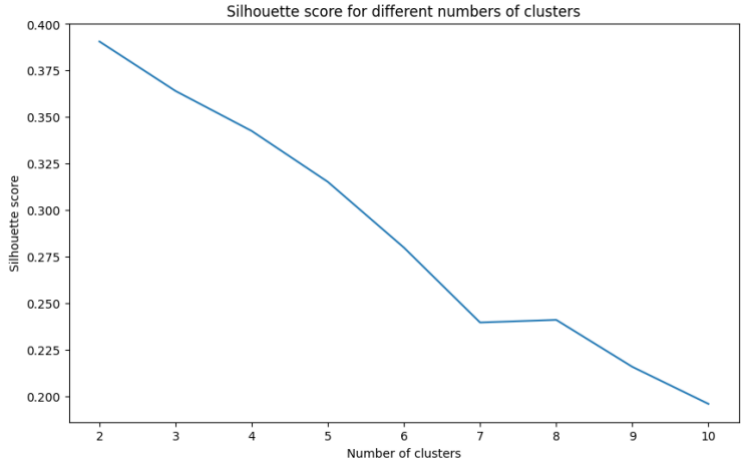
Wybrany algorytm grupowania hierarchicznego to algorytm K-medoidów.
Został on uruchomiony 4 razy, dla dwóch miar odległości – Euklidesowej i Kosinusowej oraz dla dwóch różnych ilości klastrów.

W celu optymalnego wyboru ilości klastrów utworzono wykres silhouette scores dla liczby klastrów od 2 do 10 w następujący sposób:

- Iteracja przez różne liczby klastrów od 2 do 10.
- Dla każdej liczby klastrów, utworzono instancję algorytmu K-medoidów.
- Dopasowano model.
- Obliczono wynik sylwetkowy (silhouette score) dla uzyskanych klastrów.

Proces ten przeprowadzono dwukrotnie, kolejno dla miary odległości euklidesowej, a następnie kosinusowej.



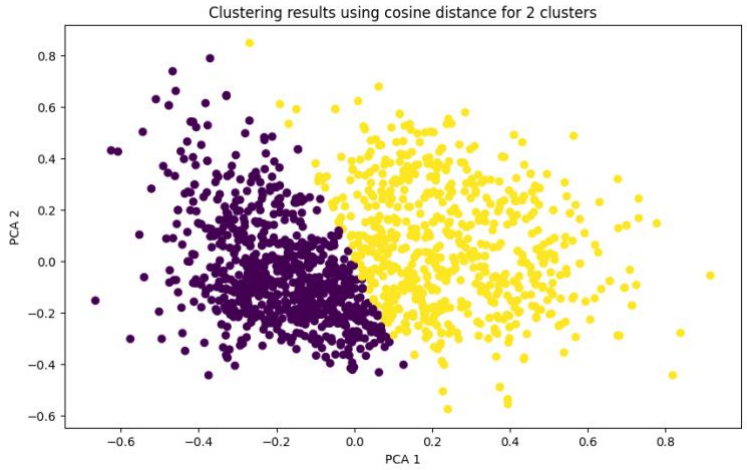


Na podstawie analizy wykresów wybrano liczby klastrów 2 oraz 5.

	Ilość klastrów i miary odległości	Indeks Daviesa-Bouldina
1	2 klastry, miara Cosine	1.052694351432217
2	5 klastrów, miara Cosine	0.9251047463576401
3	2 klastry, miara Euclidean	1.0361963522716529
4	5 klastrów, miara Euclidean	0.8658619829979081

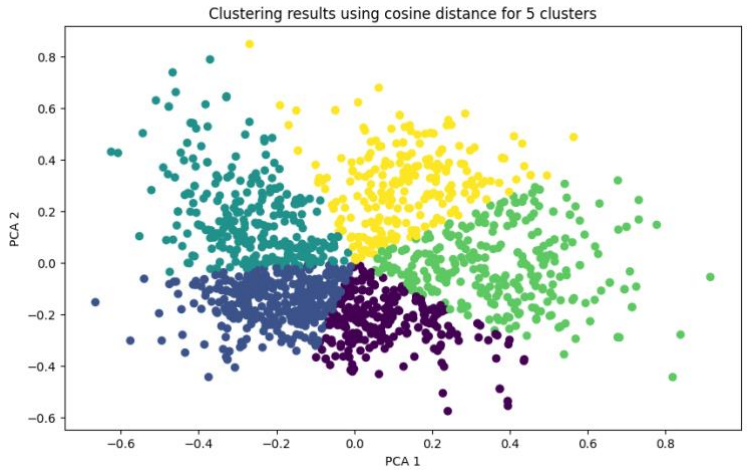
1. 2 klastry, Cosine 1.052694351432217.

Podział na 2 klastry z użyciem miary Cosine ma umiarkowaną jakość.



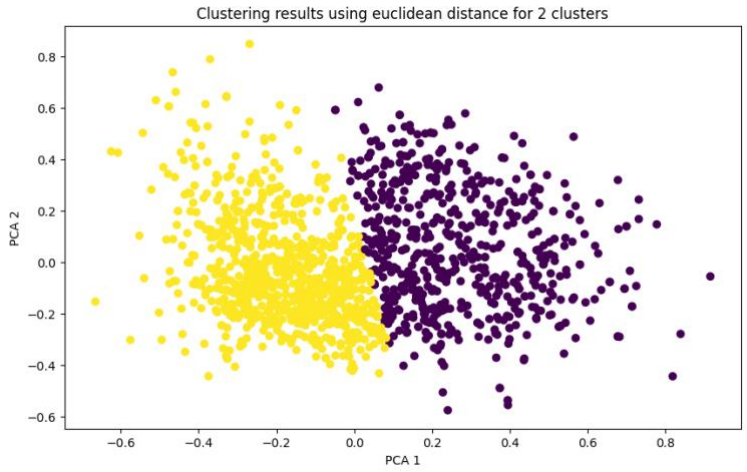
2. 5 klastrów, Cosine 0.9251047463576401.

Niższa wartość indeksu w porównaniu do przypadku z 2 klastrami sugeruje, że podział na 5 klastrów z miarą Cosine jest lepszy z punktu widzenia odległości między klastrami.



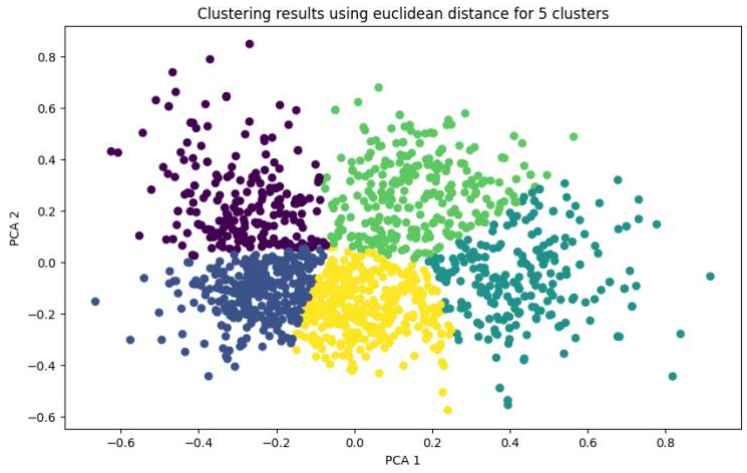
3. 2 klastry, Euclidean 1.0361963522716529.

Podobnie jak w przypadku 2 klastrów z miarą Cosine, podział na 2 klastry z miarą Euclidean ma umiarkowaną jakość.



4. 5 klastrów, Euclidean 0.8658619829979081.

Najniższa wartość indeksu wśród przedstawionych przypadków, co sugeruje, że podział na 5 klastrów z użyciem miary Euclidean jest najbardziej korzystny z punktu widzenia odległości między klastrami.



4. Grupowanie gęstościowe

Do przeprowadzenia grupowania gęstościowego wykorzystany został algorytm DBSCAN. Został on uruchomiony 4 razy, dla dwóch wartości Epsilon oraz dla dwóch różnych wartości MinPts.

Wartość epsilon to maksymalna odległość między dwiema próbkami, aby jedną można było uznać za znajdującą się w sąsiedztwie drugiej. Natomiast min_pts to minimalna liczba obiektów wchodzących w skład grupy.

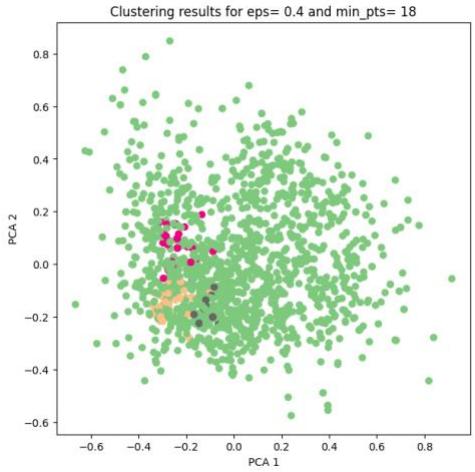
Lp.	Epsilon	min_pts	clusters	noise	davies-bouldin
1	0.400000	8	7	1222	2.143
2	0.400000	18	3	1480	1.854
3	0.500000	8	4	680	3.193
4	0.500000	18	3	994	2.787

1. Dla Epsilon 0,4 i min_pts 8, algorytm DBSCAN wyodrębnił 7 skupień i 1222 punkty odstające. Wartość wskaźnika Davies-Bouldina 2,143.

Ilość klastrów, może oznaczać, że algorytm dobrze radzi sobie z grupowaniem danych. Jednak miara Daviesa-Bouldina wynosi 2.143, co sugeruje dosyć duży stopień rozproszenia klastrów. Poza tym aż 76% zbioru zostało potraktowane jako szum.

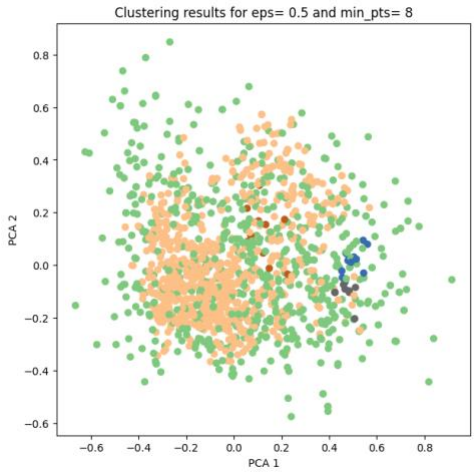
2. Dla Epsilon 0,4 i min_pts 18, algorytm DBSCAN wyodrębnił 3 skupienia i 1480 punktów odstających. Wartość wskaźnika Davies-Bouldina 1,854.

W porównaniu do poprzedniej próby, liczba klastrów spada do 3. Miara Daviesa-Bouldina jest niższa (1.854), co może wskazywać na bardziej zwarte klastry niż w poprzednim przypadku. Aż 91% zbioru zostało potraktowane jako szum.



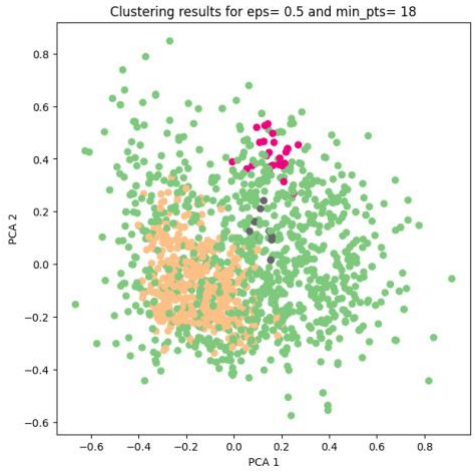
3. Dla Epsilon 0,5 i min_pts 8, algorytm DBSCAN wyodrębnił 4 skupienia i 680 punktów odstających. Wartość wskaźnika Davies-Bouldina 3,193.

Liczba klastrów wynosi 4, ale miara Daviesa-Bouldina wynosi 3.193, co sugeruje większy stopień rozproszenia i mniej jednolite klastry. 42% zbioru zostało potraktowane jako szum.



4. Dla Epsilon 0,5 i min_pts 18, algorytm DBSCAN wyodrębnił 3 skupienia i 994 punktów odstających. Wartość wskaźnika Davies-Bouldina wynosi 2,787.

Liczba klastrów wynosi 3, a miara Daviesa-Bouldina jest mniejsza niż w trzecim przypadku (2.787). To może sugerować lepszą jakość klastrów. 62% zbioru zostało potraktowane jako szum.



5. Porównanie wyników algorytmów

5.1. AHC

Na podstawie przedstawionych danych można stwierdzić, że wyniki algorytmu AHC są w ogólnym zarysie dobre. Wszystkie wartości indeksu Daviesa-Bouldina są niższe niż 2, a połowa mniejsza niż 1 co oznacza, że klastry utworzone przez algorytm AHC są dobrze zdefiniowane i odseparowane od siebie.

5.1.1. Metody łączenia

W przypadku algorytmu AHC można zauważyć, że metoda single linkage daje ogólnie lepsze wyniki niż metody complete linkage i average linkage. Oznacza to, że klastry utworzone przez metodę single linkage są bardziej jednorodne i lepiej odseparowane od siebie.

5.1.2. Miary odległości

W przypadku odległości euklidesowej wyniki algorytmu AHC są ogólnie lepsze niż w przypadku odległości kosinusowej. Oznacza to, że klastry utworzone przez algorytm AHC są bardziej jednorodne i lepiej odseparowane od siebie, gdy są mierzone za pomocą odległości euklidesowej dla tego zbioru danych.

5.2. K-medoidów

Można stwierdzić, że wyniki algorytmu k-medoidów są bardzo dobre, choć gorsze od niektórych przypadków AHC. Większość wartości indeksu Daviesa-Bouldina oscyluje wokół wartości 1, co oznacza, że klastry utworzone przez algorytm k-medoidów są dobrze zdefiniowane i odseparowane od siebie.

5.2.1. Ilość klastrów

W przypadku algorytmu k-medoidów można zauważyć, że zwiększenie ilości klastrów z 2 do 5 nieco poprawia jakość klastrów.

5.2.2. Miary odległości

W przypadku odległości euklidesowej wyniki algorytmu k-medoidów są nieznacznie lepsze niż w przypadku odległości kosinusowej.

5.3. DBSCAN

Na podstawie danych przedstawionych w tabeli można stwierdzić, że wyniki algorytmu DBSCAN są najgorsze ze wszystkich. Wartości indeksu DBI przekraczają wartość 2, co oznacza, że klastry utworzone przez algorytm DBSCAN nie są dobrze zdefiniowane, ani odseparowane od siebie.

5.3.1. Epsilon

W przypadku algorytmu DBSCAN można zauważyć, że zmniejszenie wartości epsilon z 0,5 do 0,4 poprawia jakość klastrów. Aczkolwiek, znacznie zwiększa się ilość punktów oznaczonych jako szum.

5.3.2. min_pts

W przypadku algorytmu DBSCAN można zauważyć, że zwiększenie wartości min_pts z 8 do 18 poprawiło jakość klastrów.

5.4. Podsumowanie

Na podstawie porównania wyników trzech algorytmów można stwierdzić, że najlepsze wyniki dla zbioru ‘Red wine quality’ udało się osiągnąć za pomocą algorytmu AHC dla odległości euklidesowej i metody single linkage.

Algorytm k-medoidów również daje dobre wyniki, zwłaszcza w przypadku odległości euklidesowej i 5 klastrów.

Algorytm DBSCAN dał najgorsze wyniki. Dla wielu prób większość zbioru była traktowana jako szum, a więc te punkty były nieklasyfikowane do żadnego z klastrów. W sprawozdaniu została przedstawiona wyłącznie niewielka część prób wykonanych z tym algorytmem, niestety dla żadnych wartości nie udało się dobrze rozgraniczyć klastrów. Może to wynikać z tego, że właściwości samego zbioru danych, takie jak gęstość punktów czy ich rozmieszczenie, wpłynęły na trudności zastosowania algorytmu DBSCAN.

Podsumowując, na podstawie analizy wyników trzech różnych algorytmów grupowania dla zbioru danych dotyczącego jakości czerwonych win, można wskazać, że najbardziej efektywne rezultaty uzyskano przy użyciu algorytmu AHC z odległością euklidesową i metodą pojedynczego połączenia oraz algorytmu k-medoidów, zwłaszcza przy zastosowaniu odległości euklidesowej i przy podziale na 5 klastrów.