234053
Numer indeksu
Paweł Galewicz
Imię i nazwisko

234067
Numer indeksu
Bartosz Jurczewski
Imię i nazwisko

234102
Numer indeksu
Zbigniew Nowacki
Imię i nazwisko

234106
Numer indeksu
Karol Podlewski
Imię i nazwisko

234128
Numer indeksu
Piotr Wardęcki
Imię i nazwisko

Kierunek Informatyka Stosowana
Stopień II
Specjalizacja Data Science
Semestr 1

Data oddania 15 kwietnia 2020

Metody uczenia maszynowego

Problem set 2

Spis treści

1 Cel

Zadanie polegało na analizie procesu grupowania danych za pomocą wybranych metod:

1. Algorytmu EM,
2. Algorytmu k-średnich,
3. Algorytmu hierarchiczne aglomeracyjnego,
4. Metody gęstościowej DBSCAN,
5. metoda optics.

Należało zaimplementować każdą metodę, a następnie zweryfikować jej działanie biorąc pod uwagę:

- Rożne możliwe ustawienia parametrów konfiguracyjnych i ich wpływ na wyniki klasyfikacji,
- Zbiory danych o różnej charakterystyce (przynajmniej 3 różne zbiory).

Każdą metodę należało przetestować na tych samych zbiorach, a następnie porównać wyniki i wyciągnąć wnioski dotyczące skuteczności poszczególnych metod. Jako kryterium porównawcze wykorzystaliśmy metryki *Silhouette*, *Calinski-Harabasz* oraz *Daves-Bouldin*.

2 Opis implementacji

Algorytmy zostały zaimplementowane za pomocą języka Python w wersji 3.8.2. Wykorzystano w nim biblioteki NumPy, Sklearn i Pandas.

3 Opis zbiorów danych

Bazowaliśmy na trzech zestawach danych:

A. Mall Customer Segmentation - Zawiera informacje zbierane od klientów centrum handlowego uczestniczących w programie lojalnościowym. Są to między innymi dane o wieku, płci, rocznym przychodzie. Każdy klient ma też wyliczoną ocenę wydatków w sklepie w przedziale od 1 do 100.

B. Wine quality selection – Zestaw danych przedstawiający współczynniki wina wraz z przyporządkowaną ich jakością.

C. Credit Card Dataset – Zbiór opisujący zachowania aktywnych klientów bankowych, agregowanych przez 6miesięcy.

4 Klasyfikatory

4.1 Algorytm EM

Algorytm Oczekiwania-Maksymalizacji (ang. *Expectation-Maximization*, EM) wylicza rozkłady prawdopodobieństwa przynależności obserwacji do skupień. Jego przebieg rozpoczyna się od podania liczby centrów. Dla każdego z nich losowana jest jedna obserwacja, która będzie reprezentowała dane skupienie. Następnie powtarzane są dwa kroki aż do osiągnięcia wymaganej zbieżności:

- oczekiwanie – liczenie prawdopodobieństwa przynależności obserwacji do skupień
- maksymalizacja – wyznaczenie wartości parametrów skupień, przy których wiarygodność rozkładu jest maksymalna

Parametry, które postanowiono przetestować jest liczba maksymalnych iteracji algorytmu oraz typ macierzy kowariancji.

4.2 Algorytm k-średnich

Algorytm k-średnich dzieli zbiór przypadków na k skupień, gdzie k jest podaną wartością. Rozpoczynając działanie od losowo wybranych środków skupień, przypisuje do nich obiekty biorąc pod uwagę odległość obiektów od wybranych środków. Kiedy obiekty zostały przypisane, wyznaczane są nowe środki, które zostają wyliczone na podstawie wartości atrybutów obiektów należących do danej grupy skupień.

Algorytm wymaga wcześniejszego określenia liczby skupień oraz maksymalnej liczby iteracji, po której przestanie dalej przetwarzać dane jeżeli wcześniej nie osiągnął stabilizacji. W przypadku zaprezentowanego rozwiązania użytkownik ma też wpływ na początkowe rozłożenie centrów oraz ilość przeprowadzonych losowań centrów w obrębie jednego uruchomienia algorytmu.

4.3 Algorytm hierarchiczny aglomeracyjny

Metody hierarchiczne aglomeracyjne na początku traktują wszystkie obserwacje jak osobne klastry. W kolejnych krokach najbardziej zbliżone do siebie pary klastrów są ze sobą łączone - poruszamy się w górę hierarchii. Dostępne parametry to:

- metryka
  - euklidesowa $d(p,q) = \sqrt{\sum_i (q_i - p_i)^2}$
  - manhattan $d(p,q) = \sum_i |q_i - p_i|$
- metoda łączenia
  - Warda – odległość między klastrami jest sumą kwadratów odchyleń od punktów do centroidów. Dąży do zminimalizowania sumy kwadratów wewnątrz klastra
  - kompletne – odgległość między klastrami jest maksymalną odległością między obserwacją w jednym i drugim klastrze. Wrażliwy na obserwacje odstające od pozostałych w próbie.
  - średnie – odległość między klastrami jest średnią odległością między obserwacjami w jednym i drugim klastrze
  - pojedyncze – odległość między klastrami jest minimalną odległością między obserwacjami w jednym i drugim klastrze. Najlepsze kiedy klastry są wyraźnie oddzielone

Parametry, które postanowiono przetestować to obie metryki oraz metody połączeń pojedynczych i średnich.

4.4 Metoda gęstościowa DBSCAN

Algorytm klasteryzcji opracowany pod koniec lat 90 ubiegłego wieku. Jego klasteryzacja opiera się na gęstości punktów w skupieniach. Przyjmując na wejściu zbiór punktów, grupuje razem te punkty które są najbliższymi sąsiadami (najbliżej "upakowane"). Należy do grona najpopularniejszych algorytmów i jest najczęściej wspominanym w naukowej literaturze.

4.5 Algorytm optics

Algorytm ten jest algorytmem w oparciu o znalezienie gęstości klastrów w danych przestrzennych. Jest algorytmem zbliżonym do algorytmu DBSCAN jednak z pewnymi przewagami, potrafi wykrywać znaczące klastry w danych o różnej gęstości. W tym celu punkty danych są uporządkowane liniowo w taki sposób, że najbliższe przestrzennie punkty stają się sąsiadami. Dla każdego punktu zapisywana jest specjalna odległość reprezentująca gęstość oraz określająca przynależność do konkretnego zbioru.

5 Badania

W sekcji badań odnoszono się do zbiorów danych za pomocą oznaczeń zgodnych z sekcją 3. Przy badaniu wykorzystano następujące metryki:

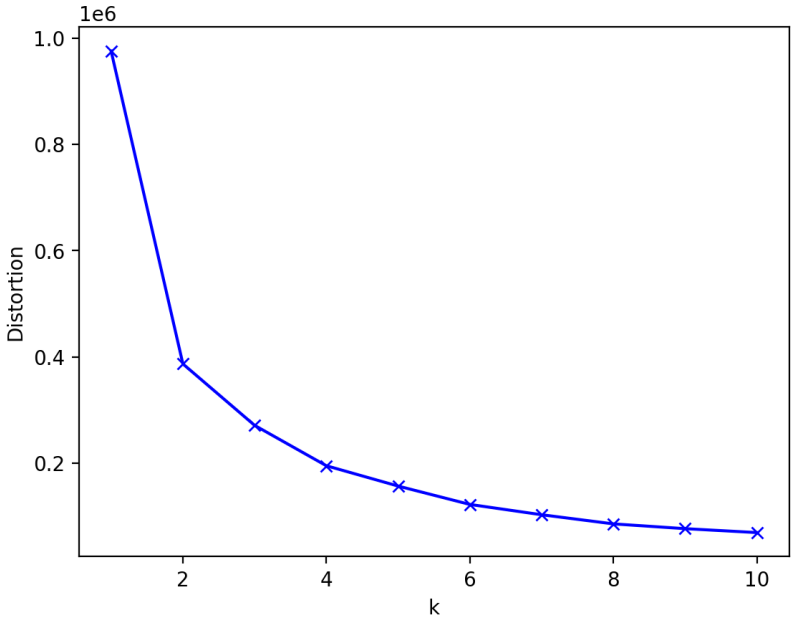
1. Indeks Silhouette,
2. Indeks Calińskiego-Harabasza,
3. Indeks Daviesa-Bouldina.

Żadna z metryk nie odnosi się do oczekiwanych rezultatów. Dla każdej z nich lepszym wynikiem będzie wyższa liczba. Nie można jednak porównywać metryk między sobą.

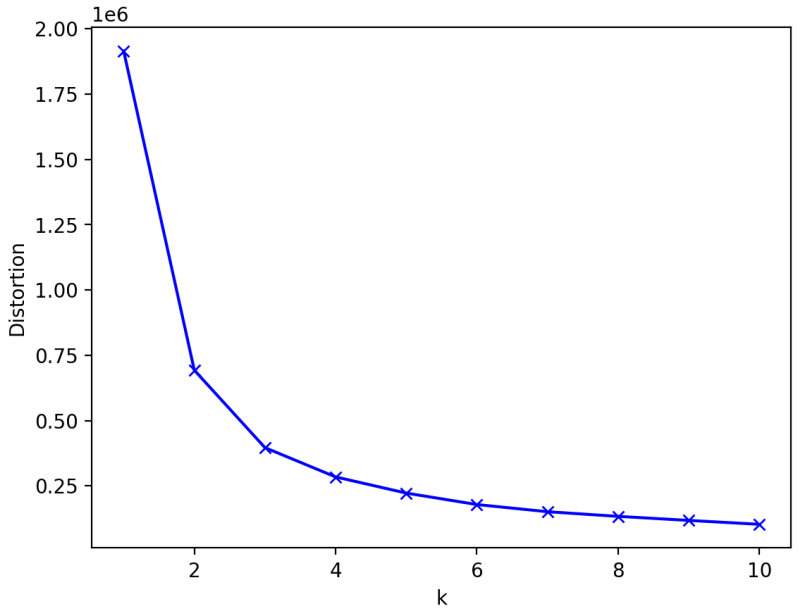
Porównywanie uzyskanych rezultatów pomiędzy algorytmami także nie jest wysoce miarodajne - metryki mogą oczekiwać przypasowania danych w określony sposób, niezależnie od użytego algorytmu. Będzie to promować rozwiązania, które są bliżej oczekiwanego wzorca, nie zawsze będąc faktycznie optymalnym podziałem dla danego zbioru danych [1]. Dlatego zdecydowaliśmy się nie porównywać bezpośrednio uzyskanych wyników pomiędzy różnymi algorytmami.

5.1 Liczba centrów dla zbiorów danych

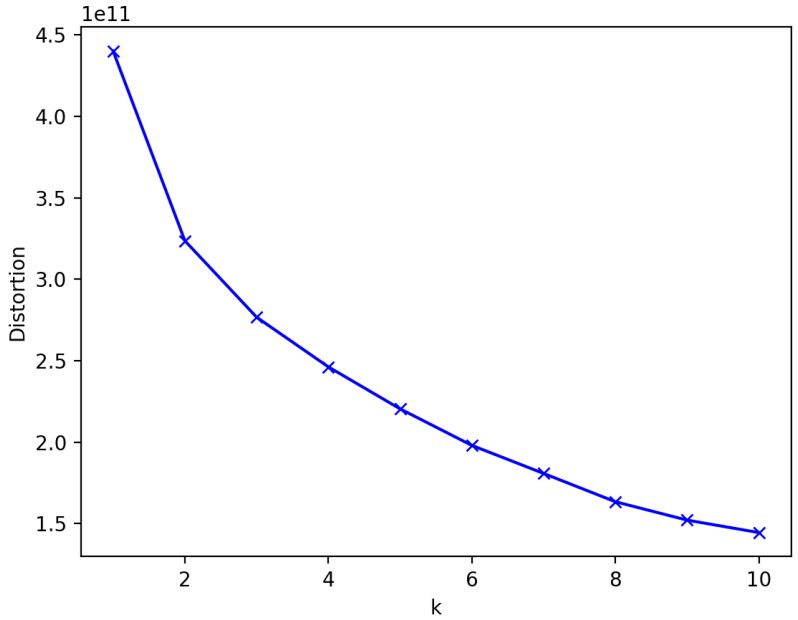
W celu odpowiedniego doboru liczby centrów dla algorytmów Oczekiwania-Maksymalizacji, k-średnich oraz hierarchicznie aglomeracyjnego przeprowadzano testy wykorzystując metodę "łokcia" (która wykorzystuje algorytm k-średnich).



Rysunek 1: Metoda "łokcia" dla zbioru danych A



Rysunek 2: Metoda ”łokcia” dla zbioru danych B



Rysunek 3: Metoda ”łokcia” dla zbioru danych C

Wyniki uzyskane z wykorzystaniem metody ”łokcia” dla zbioru danych A sugerowały użycie 4 lub 5 klastrów - my zdecydowaliśmy się na ten drugi wariant. W przypadku danych B były to 4 klastry, a w przypadku danych C - 3 klastry.

5.2 Algorytm EM

Zbiór A

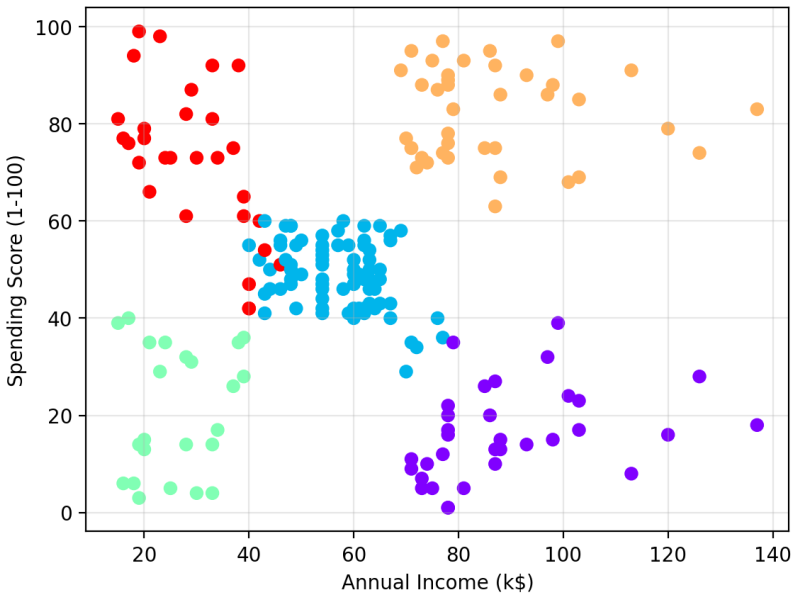
Typ kowariancji	Silhouette	Calinski-Harabasz	Davies-Bouldin
full	0.35	223.102	0.905
tied	0.398	212.722	0.898
diag	0.343	231.842	0.989
spherical	0.442	251.828	0.79

Tabela 1: Wartości metryk dla różnych typów macierzy kowariancji algorytmu EM dla zbioru A

Iteracja	Silhouette	Calinski-Harabasz	Davies-Bouldin
20	0.379	229.571	0.893
40	0.351	214.546	0.974
60	0.379	229.571	0.893
80	0.351	214.546	0.974
100	0.444	243.475	0.737
120	0.313	163.594	1.176
140	0.397	211.897	0.899
160	0.407	247.42	0.893
180	0.33	197.16	0.954
200	0.351	214.546	0.974

Tabela 2: Wartości metryk dla różnej maksymalnej liczby iteracji algorytmu EM dla zbioru A

Na podstawie tabel 1 oraz 2 wywnioskowano, że najkorzystniejsze wyniki wychodzą dla typu kowariancji *"full"* oraz maksymalnej liczbie iteracji na poziomie 160. Konfigurację tę wykorzystano do wizualizacji danych na rysunku 4.



Rysunek 4: Wizualizacja wyliczonych centrów dla zbioru A przy pomocy EM

Zbiór B

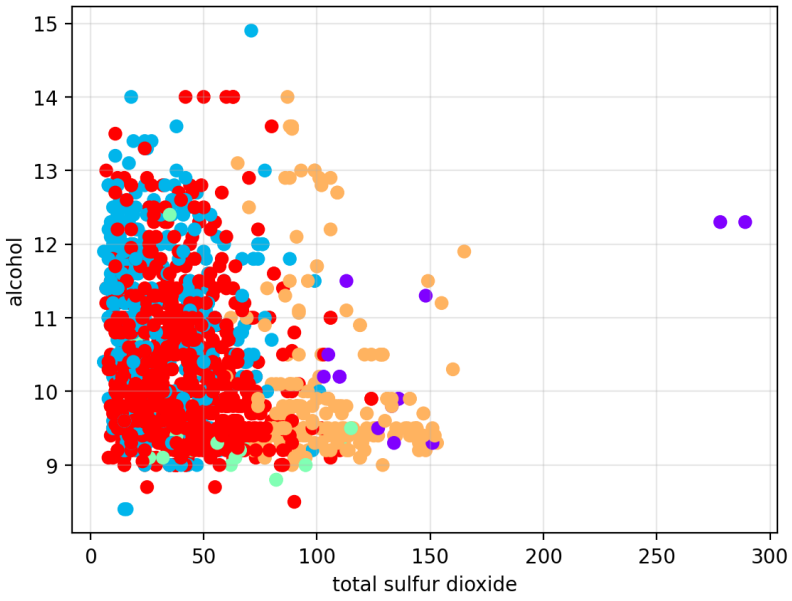
Typ kowariancji	Silhouette	Calinski-Harabasz	Davies-Bouldin
full	-0.165	82.86	4.17
tied	0.286	702.037	1.273
diag	-0.162	101.208	6.883
spherical	0.368	2116.932	0.815

Tabela 3: Wartości metryk dla różnych typów macierzy kowariancji algorytmu EM dla zbioru B

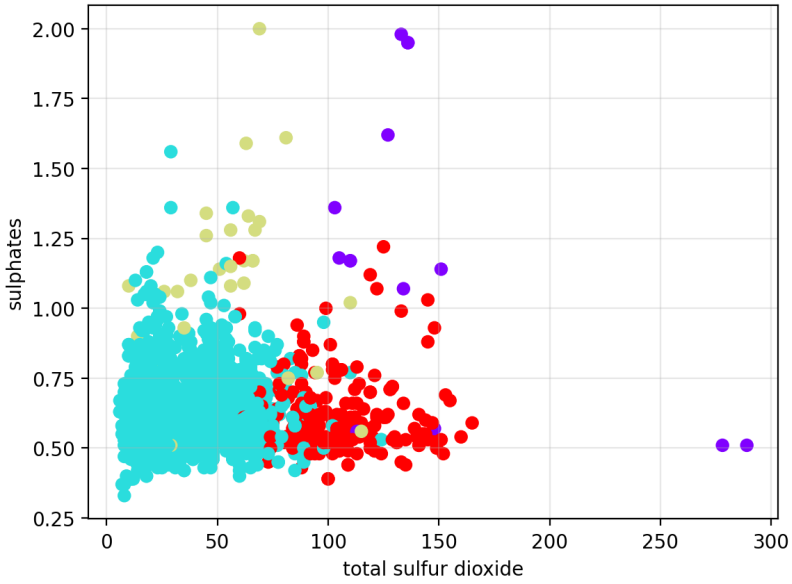
Iteracja	Silhouette	Calinski-Harabasz	Davies-Bouldin
20	0.048	338.056	3.437
40	-0.157	124.091	7.692
60	-0.165	82.86	4.17
80	-0.156	124.344	7.692
100	-0.004	77.789	10.078
120	-0.11	93.36	5.182
140	-0.117	98.011	6.9
160	-0.156	124.344	7.692
180	-0.105	91.154	5.671
200	-0.156	124.344	7.692

Tabela 4: Wartości metryk dla różnej maksymalnej liczby iteracji algorytmu EM dla zbioru B

Najlepszą konfiguracją dla zbioru B jest *tied* oraz 100 iteracji. Na jej podstawie wygenerowano wykresy.



Rysunek 5: Wizualizacja centrów dla zbioru B przy pomocy EM



Rysunek 6: Wizualizacja centrów dla zbioru B przy pomocy EM

Zbiór C

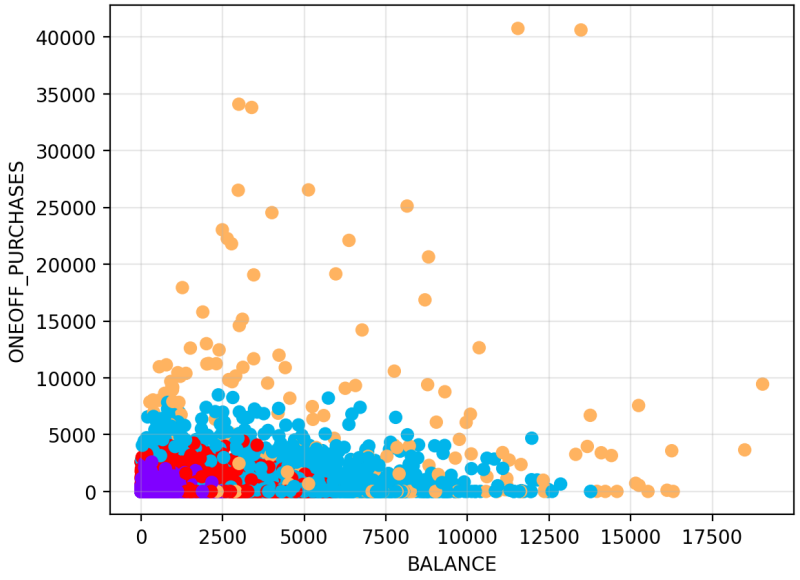
Typ kowariancji	Silhouette	Calinski-Harabasz	Davies-Bouldin
full	0.014	359.952	5.903
tied	0.194	1258.45	1.742
diag	-0.064	440.984	4.07
spherical	0.204	1427.614	1.499

Tabela 5: Wartości metryk dla różnych typów macierzy kowariancji algorytmu EM dla zbioru C

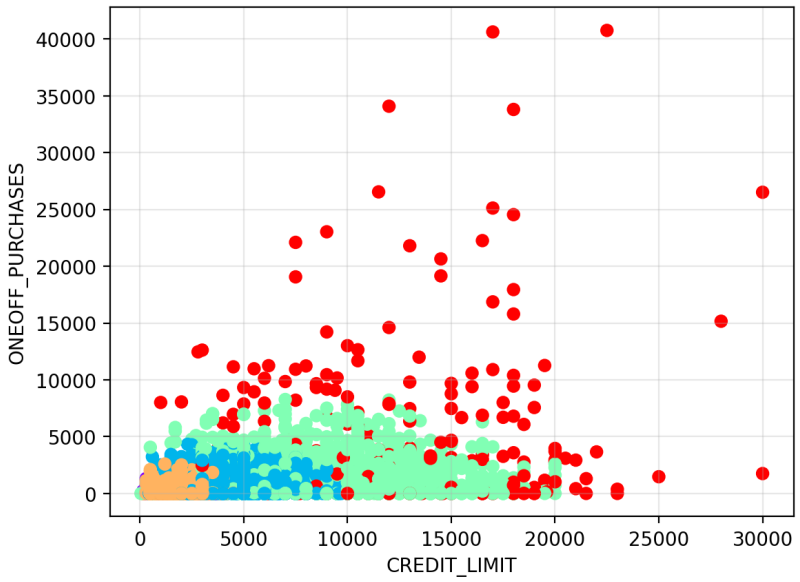
Iteracja	Silhouette	Calinski-Harabasz	Davies-Bouldin
20	0.015	400.416	3.93
40	-0.082	442.734	3.262
60	-0.038	278.84	4.683
80	0.014	361.875	5.897
100	-0.082	443.708	3.259
120	-0.079	446.098	3.306
140	0.014	359.952	5.903
160	-0.033	443.15	5.37
180	-0.079	350.615	4.564
200	-0.033	443.15	5.37

Tabela 6: Wartości metryk dla różnej maksymalnej liczby iteracji algorytmu EM dla zbioru C

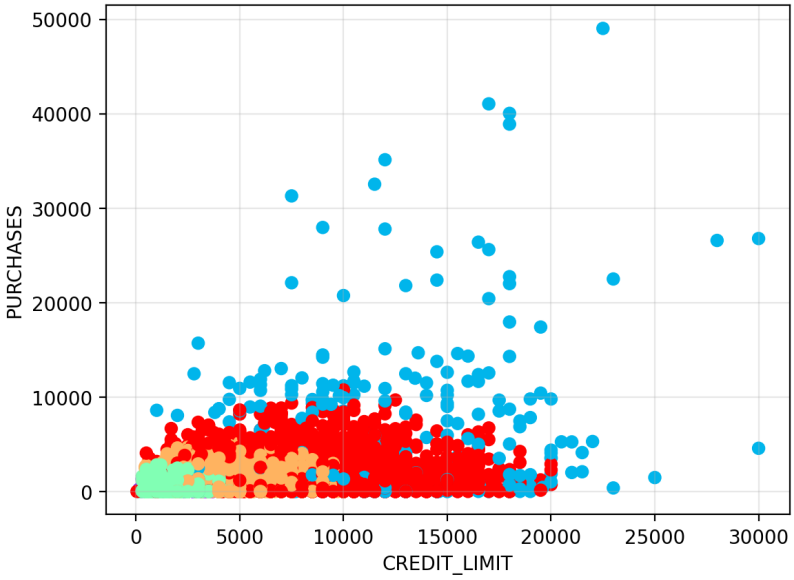
Z tabel 5 oraz 6 wywnioskować można, że najlepszą konfiguracją jest *spherical* oraz 100 iteracji.



Rysunek 7: Wizualizacja centrów dla zbioru C przy pomocy EM



Rysunek 8: Wizualizacja centrów dla zbioru C przy pomocy EM



Rysunek 9: Wizualizacja centrów dla zbioru C przy pomocy EM

5.3 Algorytm k-średnich

Na samym początku sprawdzono wpływ powtórnego losowania centrów na uzyskane wyniki. Następnie porównano liczbę maksymalnych iteracji, po czym sprawdzono wpływ wyboru ziarna przy losowaniu centrów. Ziarna zostały wybrane zgodnie z [2].

Standardowa konfiguracja obejmowała 10 uruchomień z losowaniem centrów, 50 maksymalnych iteracji oraz brak zmiany ziarna losowania.

Zbiór A

Uruchomienia	Silhouette	Calinski-Harabasz	Davies-Bouldin
1	0.3497	205.5867	0.9468
2	0.4204	253.8026	0.8705
5	0.4400	252.5983	0.8046
10	0.4251	253.8095	0.8532
20	0.4247	253.8266	0.8568

Tabela 7: Wartości metryk dla różnej liczby losowań początkowych centrów, dla zbioru A z wykorzystaniem algorytmu k-średnich

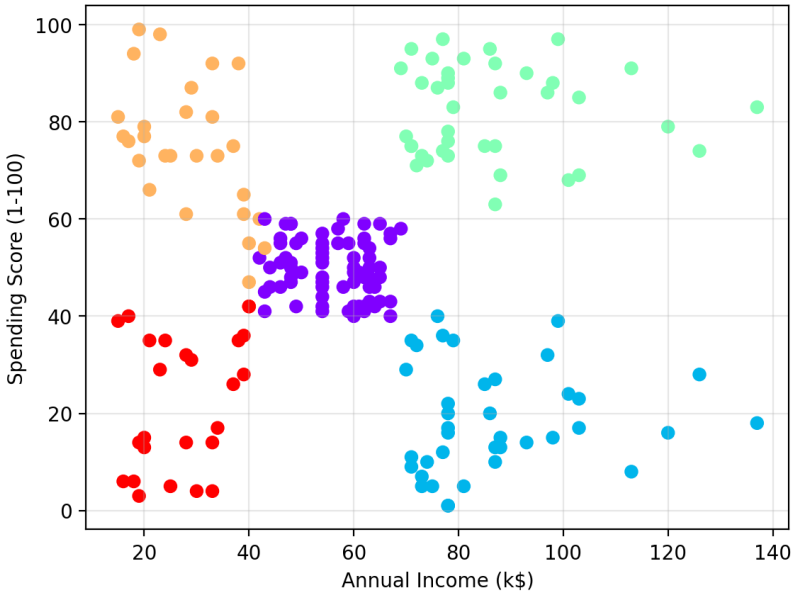
Maks. iteracje	Silhouette	Calinski-Harabasz	Davies-Bouldin
5	0.4400	252.5983	0.8046
10	0.4247	253.8266	0.8568
20	0.4204	253.8026	0.8705
50	0.4337	253.1832	0.8343
100	0.4337	253.1832	0.8343

Tabela 8: Wartości metryk dla różnej liczby maksymalnych iteracji, dla zbioru A z wykorzystaniem algorytmu k-średnich

Ziarno	Silhouette	Calinski-Harabasz	Davies-Bouldin
Domyślne	0.4332	253.0810	0.8233
0	0.4282	253.8041	0.8456
42	0.4216	253.8315	0.8668

Tabela 9: Wartości metryk dla różnego ziarna, dla zbioru A z wykorzystaniem algorytmu k-średnich

Uzyskane wyniki wskazują na brak większej zależności zmiany parametrów na uzyskane rezultaty wybranych metryk. Na wizualizacji 10 zastosowano domyślne parametry uruchomienia.



Rysunek 10: Wizualizacja centrów dla zbioru A z wykorzystaniem algorytmu k-średnich

Zbiór B

Uruchomienia	Silhouette	Calinski-Harabasz	Davies-Bouldin
1	0.4838	3051.5684	0.7150
2	0.4838	3051.5684	0.7150
5	0.4838	3051.5684	0.7150
10	0.4838	3051.5684	0.7150
20	0.4838	3051.5684	0.7150

Tabela 10: Wartości metryk dla różnej liczby losowań początkowych centrów, dla zbioru B z wykorzystaniem algorytmu k-średnich

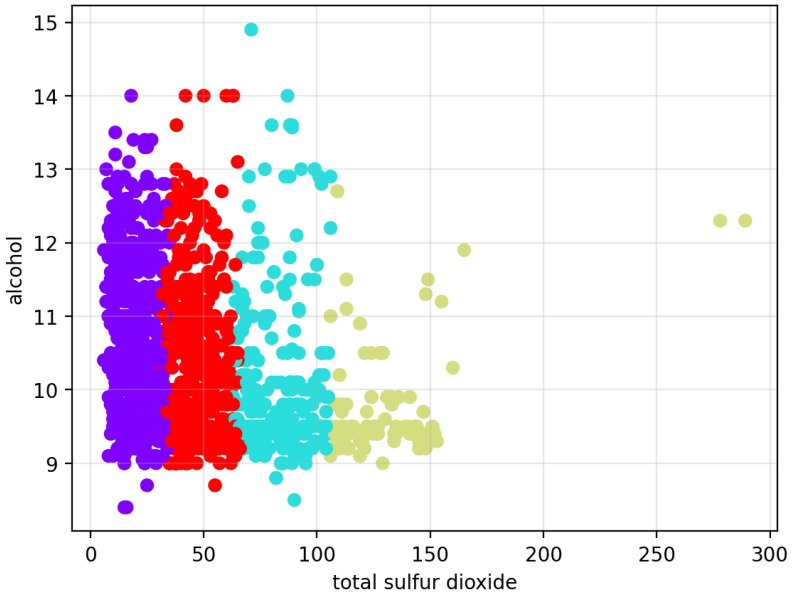
Maks. iteracje	Silhouette	Calinski-Harabasz	Davies-Bouldin
5	0.4884	3050.9843	0.7109
10	0.4833	3051.0148	0.7147
20	0.4838	3051.5684	0.7150
50	0.4838	3051.5684	0.7150
100	0.4838	3051.5684	0.7150

Tabela 11: Wartości metryk dla różnej liczby maksymalnych iteracji, dla zbioru B z wykorzystaniem algorytmu k-średnich

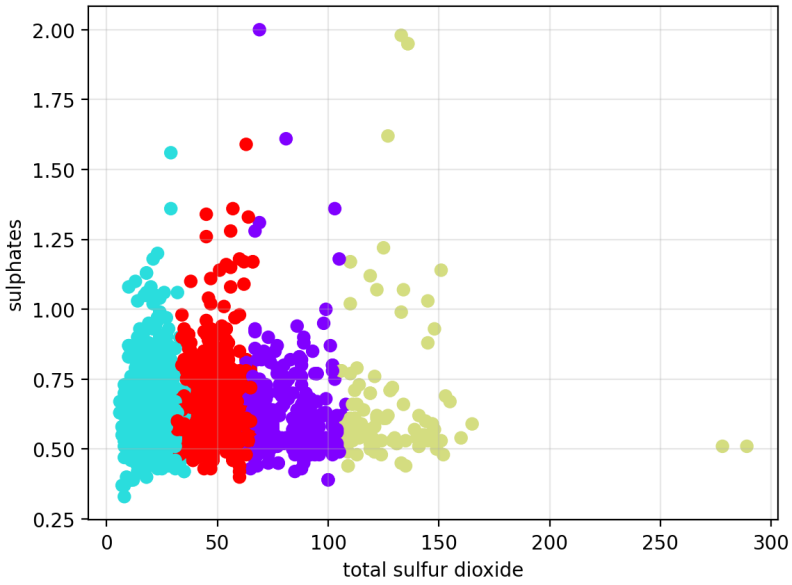
Ziarno	Silhouette	Calinski-Harabasz	Davies-Bouldin
Domyślne	0.4845	3051.9139	0.7146
0	0.4838	3051.5684	0.7150
42	0.4838	3051.5684	0.7150

Tabela 12: Wartości metryk dla różnego ziarna, dla zbioru B z wykorzystaniem algorytmu k-średnich

Uzyskane wyniki wskazują na brak większej zależności zmiany parametrów na uzyskane rezultaty wybranych metryk. Na wizualizacjach 11 oraz 12 zastosowano domyślne parametry uruchomienia.



Rysunek 11: Pierwsza wizualizacja centrów dla zbioru B z wykorzystaniem algorytmu k-średnich



Rysunek 12: Druga wizualizacja centrów dla zbioru B z wykorzystaniem algorytmu k-średnich

Zbiór C

Uruchomienia	Silhouette	Calinski-Harabasz	Davies-Bouldin
1	0.3808	2549.0832	1.2981
2	0.3808	2549.0832	1.2981
5	0.3808	2549.0832	1.2981
10	0.3820	2549.3077	1.2934
20	0.3819	2549.3082	1.2941

Tabela 13: Wartości metryk dla różnej liczby losowań początkowych centrów, dla zbioru C z wykorzystaniem algorytmu k-średnich

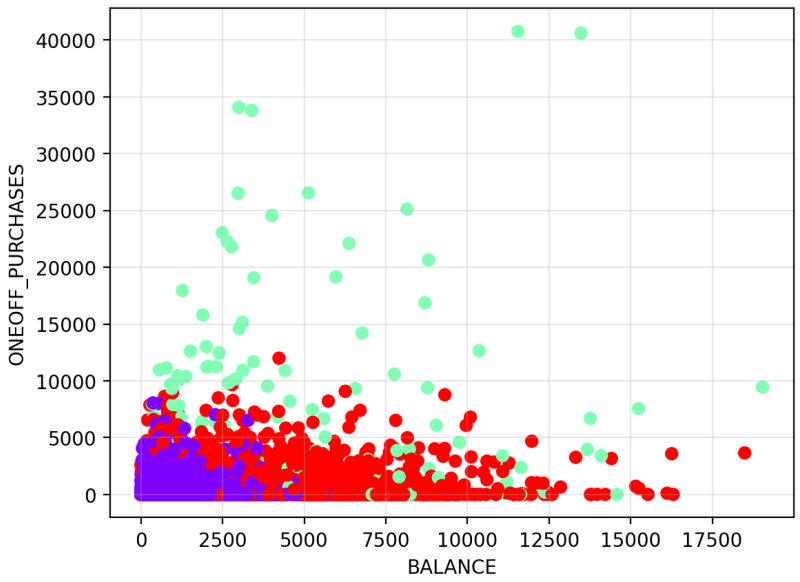
Maks. iteracje	Silhouette	Calinski-Harabasz	Davies-Bouldin
5	0.3701	2540.3679	1.3224
10	0.3748	2542.0966	1.3187
20	0.3820	2549.3077	1.2934
50	0.3808	2549.0832	1.2981
100	0.3808	2549.0832	1.2981

Tabela 14: Wartości metryk dla różnej liczby maksymalnych iteracji, dla zbioru C z wykorzystaniem algorytmu k-średnich

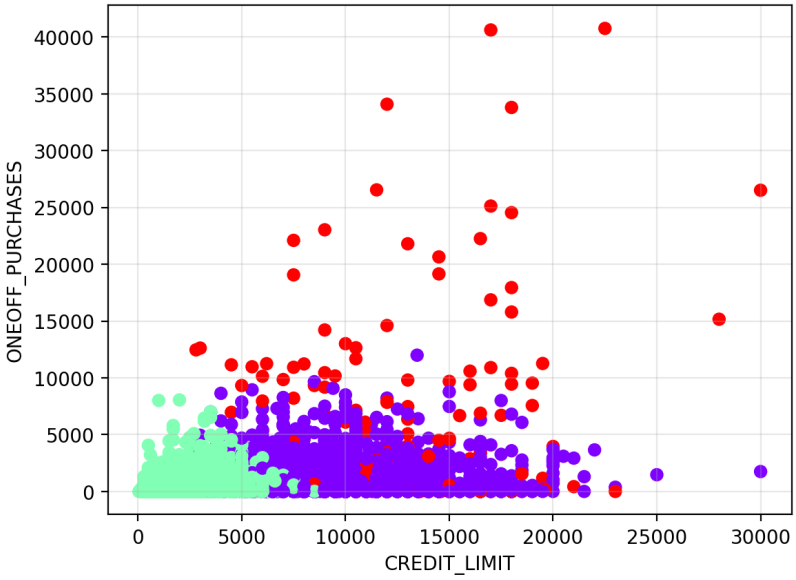
Ziarno	Silhouette	Calinski-Harabasz	Davies-Bouldin
Domyślne	0.3820	2549.3077	1.2934
0	0.3808	2549.0832	1.2981
42	0.3820	2549.3077	1.2934

Tabela 15: Wartości metryk dla różnego ziarna, dla zbioru C z wykorzystaniem algorytmu k-średnich

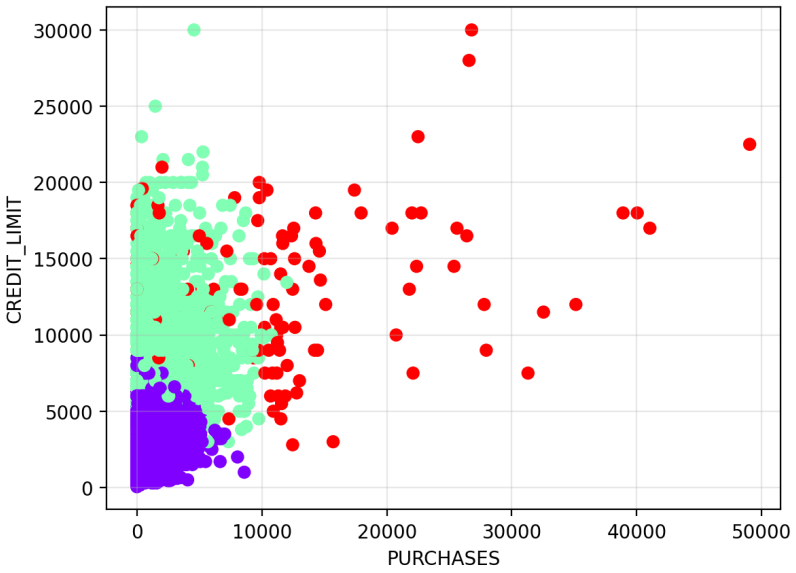
Uzyskane wyniki wskazują na brak większej zależności zmiany parametrów na uzyskane rezultaty wybranych metryk. Na wizualizacjach 13, 14 i 15 zastosowano domyślne parametry uruchomienia.



Rysunek 13: Pierwsza wizualizacja centrów dla zbioru C z wykorzystaniem algorytmu k-średnich



Rysunek 14: Druga wizualizacja centrów dla zbioru C z wykorzystaniem algorytmu k-średnich



Rysunek 15: Trzecia wizualizacja centrów dla zbioru C z wykorzystaniem algorytmu k-średnich

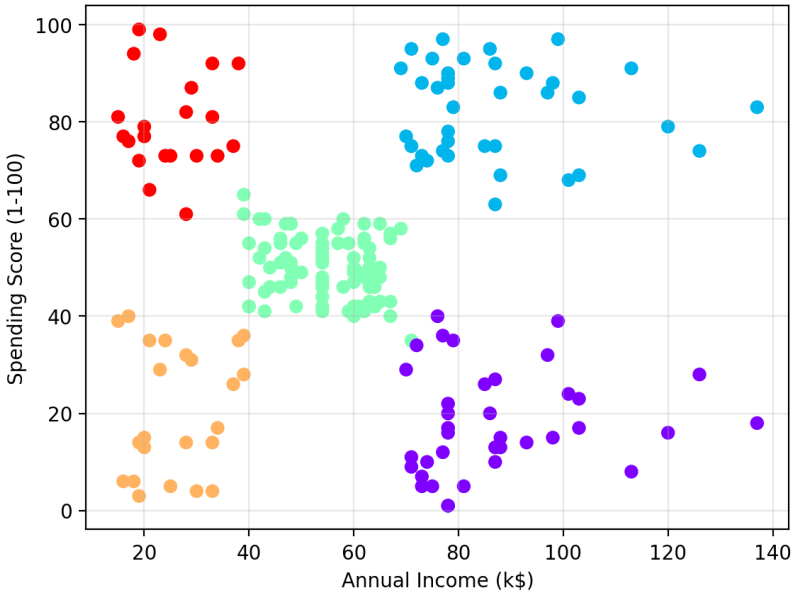
5.4 Algorytm hierarchiczne aglomeracyjny

Zbiór A

Konfiguracja	Silhouette	Calinski-Harabasz	Davies-Bouldin
single euclidean	0.212	94.893	0.693
single manhattan	0.188	17.934	0.661
average euclidean	0.443	243.445	0.741
average manhattan	0.443	243.445	0.741

Tabela 16: Wartości metryk dla różnych konfiguracji algorytmu aglomeracyjnego dla zbioru A

Na podstawie tabeli 16 wywnioskowano, że najlepszą konfiguracją jest metryka euklidesowa i połączenia średnie. Wykorzystaną ją do wizualizacji danych na rysunku 16



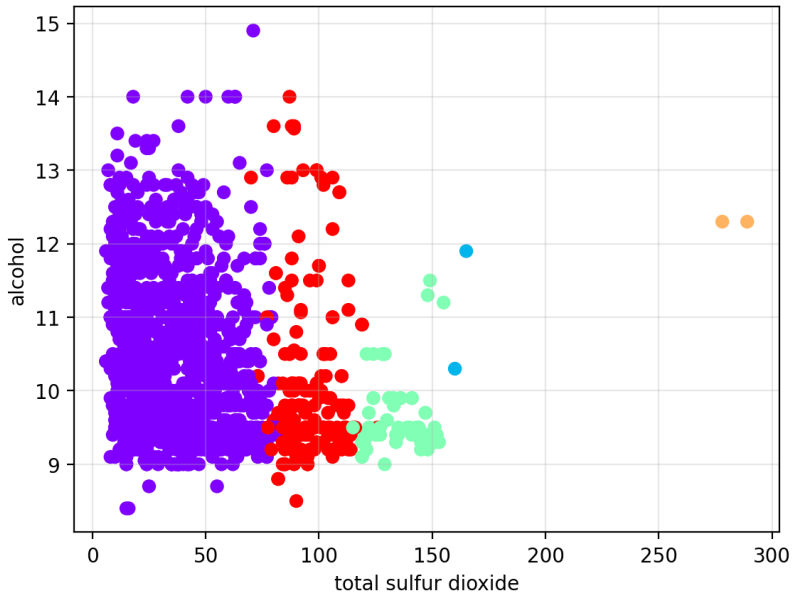
Rysunek 16: Wizualizacja centrów dla zbioru A z wykorzystaniem algorytmu aglomeracyjnego

Zbiór B

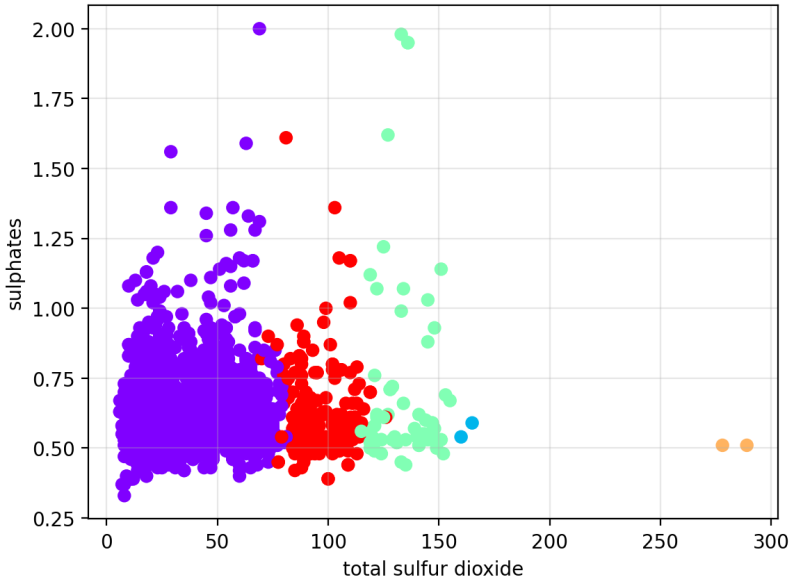
Konfiguracja	Silhouette	Calinski-Harabasz	Davies-Bouldin
single euclidean	0.405	36.180	0.298
single manhattan	0.405	36.180	0.298
average euclidean	0.534	815.003	0.604
average manhattan	0.573	753.161	0.477

Tabela 17: Wartości metryk dla różnych konfiguracji algorytmu aglomeracyjnego dla zbioru B

Na podstawie tabeli 17 widać, że ponownie najlepszą konfiguracją okazało się zestawienie metryki euklidesowej z połączeniami średnimi. Na jej podstawie wygenerowano wykresy 17 i 18.



Rysunek 17: Wizualizacja centrów dla zbioru B z wykorzystaniem algorytmu aglomeracyjnego



Rysunek 18: Wizualizacja centrów dla zbioru B z wykorzystaniem algorytmu aglomeracyjnego

Zbiór C

Badania dla zbioru C były niemożliwe do przeprowadzenia dla niniejszej metody klasteryzacji, ponieważ program potrzebował zbyt dużej ilości pamięci.

5.5 Metoda gęstościowa DBSCAN

Na początku sprawdzono jak zmiana parametru *eps* wpłynęła na uzyskane wyniki. *Eps* to maksymalny dystans między dwoma próbkami aby jedna została zakwalifikowana jako sąsiad tej drugiej.
Następnie dla ustalonego parametru sprawdzono wpływ parametru *min_samples* na wyniki. *Min_samples* to liczba próbek w sąsiedztwie dla punktu aby uznać ją za punkt rdzenny (ang. core point). Ostatnim krokiem była wizualizacja danych dla zwycięskich parametrów.

Zbiór A

eps	Silhouette	Calinski-Harabasz	Davies-Bouldin	Klastry
8	-0.2762	0.1294	5.9948	1
9	-0.3041	1.8672	3.3082	3
10	-0.2023	5.5955	2.7207	10
11	-0.1596	8.0735	2.8691	9
12	-0.0988	12.0774	2.1626	8
13	0.0312	27.5645	2.0105	4
14	0.0423	27.5407	1.5976	4
15	0.1229	36.3278	1.6726	4
16	0.0977	37.3811	1.5397	5
17	0.3386	75.8489	1.6961	3
18	0.3727	83.3958	2.0948	3
19	0.2236	32.6268	2.6009	2
20	0.2159	30.0229	3.2471	2

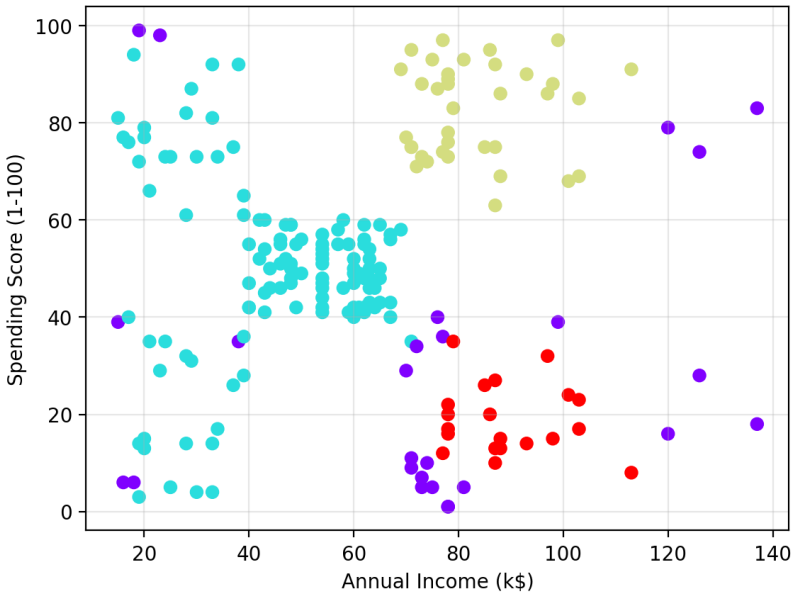
Tabela 18: Wartości metryk dla różnej wartości parametru *eps*

Uzyskane wyniki pokazują najlepsze wyniki dla *eps* równego 17.

min_samples	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**	**Klastry**
1	-0.0547	49.319	0.903	9
2	0.2044	52.1406	2.474	7
3	0.2186	59.2279	2.3552	6
4	0.2652	58.1151	2.1279	4
5	0.3386	75.8489	1.6961	3
6	0.118	40.9146	1.5171	5
7	0.1866	44.0961	1.747	5
8	0.1806	37.6648	2.3174	4
9	0.0745	26.6449	2.0137	4
10	-0.0076	21.2975	2.0046	4

Tabela 19: Wartości metryk dla różnej liczby *min_samples*

Uzyskane wyniki pokazują najlepsze wyniki dla *min_samples* równego 5.



Rysunek 19: Wizualizacja wyliczonych centrów dla zbioru A przy pomocy DBSCAN

Wizualizacja dla zwycięskich wartości: $eps = 17$, $min\_samples = 5$.

Zbiór B

eps	Silhouette	Calinski-Harabasz	Davies-Bouldin	Klastry
5	0.2817	97.6543	3.7493	6
6	0.2803	96.1985	1.0813	4
7	0.2761	82.1759	1.1353	2
8	0.5727	151.1773	0.8096	1
9	0.5842	142.9341	0.8203	1
10	0.628	133.8224	0.7671	1
11	0.6392	150.4493	0.6626	1
12	0.6682	139.4771	0.6457	1
13	0.6897	106.2034	0.7371	1
14	0.7136	107.8022	0.6816	1
15	0.7136	107.8022	0.6816	1
16	0.7136	107.8022	0.6816	1
17	0.7943	105.602	0.4244	1
18	0.7943	105.602	0.4244	1
19	0.7943	105.602	0.4244	1
20	0.7943	105.602	0.4244	1

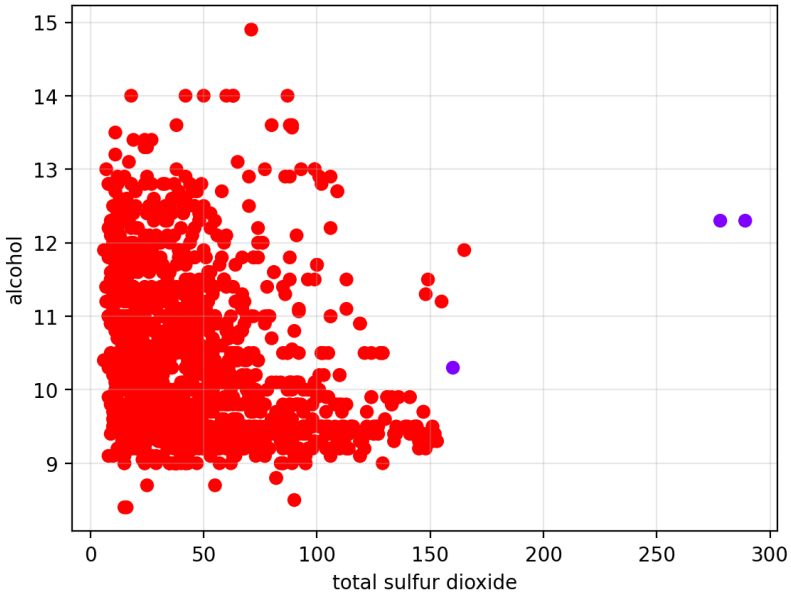
Tabela 20: Wartości metryk dla różnej wartości parametru *eps*

Uzyskane wyniki pokazują najlepsze wyniki dla *eps* równego 17 wzwyż.

min_samples	Silhouette	Calinski-Harabasz	Davies-Bouldin	Klastry
1	0.6576	58.0027	0.1897	3
2	0.6576	58.0027	0.1897	2
3	0.7943	105.602	0.4244	1
4	0.7943	105.602	0.4244	1
5	0.7943	105.602	0.4244	1
6	0.7943	105.602	0.4244	1
7	0.7943	105.602	0.4244	1
8	0.7355	106.0016	0.647	1
9	0.7355	106.0016	0.647	1
10	0.7355	106.0016	0.647	1

Tabela 21: Wartości metryk dla różnej liczby *min_samples*

Uzyskane wyniki pokazują najlepsze wyniki dla *min_samples* równego 3 (aż do 7).



Rysunek 20: Wizualizacja wyliczonych centrów dla zbioru A przy pomocy DBSCAN

Wizualizacja dla zwycięskich wartości: $eps = 17$, $min\_samples = 3$.

Zbiór C

Badania dla zbioru C były niemożliwe do przeprowadzenia dla niniejszej metody klasteryzacji, ponieważ program potrzebował zbyt dużej ilości pamięci.

5.6 Algorytm optics

Pierwszym krokiem do sprawdzenia algorytmu było ustawienie początkowych wartości dla parametrów: min-samples oraz min-cluster-size. Min-samples określa liczbę próbek leżących w sąsiedztwie punktu który bierzemy za punkt centralny natomiast min-cluster-size jest minimalna liczbą próbek w klastrze. Kolejnym korkiem było sprawdzenie otrzymanych wartości dla wybranych parametrów oraz zwizualizowanie najlepszych otrzymanych wartości.

Zbiór A

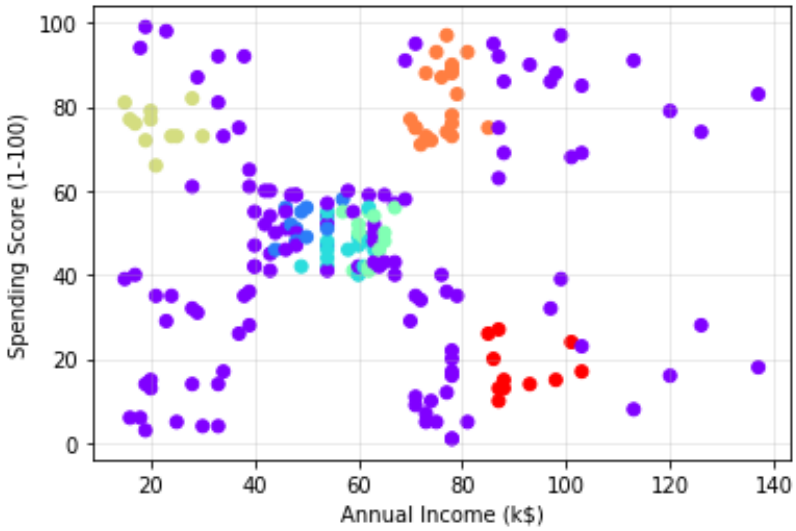
min-sample	Silhouette	Calinski-Harabasz	Davies-Bouldin
5	0.2786	44.3790	1.3174
10	0.2841	41.6443	1.5186
20	0.0455	18.0750	2.5740
30	0.1370	7.2539	4.1196
40	0.2406	11.6215	3.8548
50	0.0824	2.5170	6.6239

Tabela 22: Wartości metryk dla różnej liczby minimalnych próbek, dla zbioru A z wykorzystaniem algorytmu optics

min-cluster-size	Silhouette	Calinski-Harabasz	Davies-Bouldin
0.05	-0.1135	13.3816	3.4582
0.1	-0.0539	22.4077	2.7230
0.5	0.2786	44.3790	1.3174
2	-0.1550	8.4379	1.9298
5	-0.1793	8.9151	2.1548
10	-0.1135	13.3816	3.4582

Tabela 23: Wartości metryk dla różnej liczby minimalnych próbek, dla zbioru A z wykorzystaniem algorytmu optics

Najlepszą konfiguracją jest minimalny rozmiar klastra = 0.5 oraz minimalna liczba próbek = 5



Rysunek 21: Wizualizacja wyliczonych centrów dla zbioru A przy pomocy optics

Zbiór B

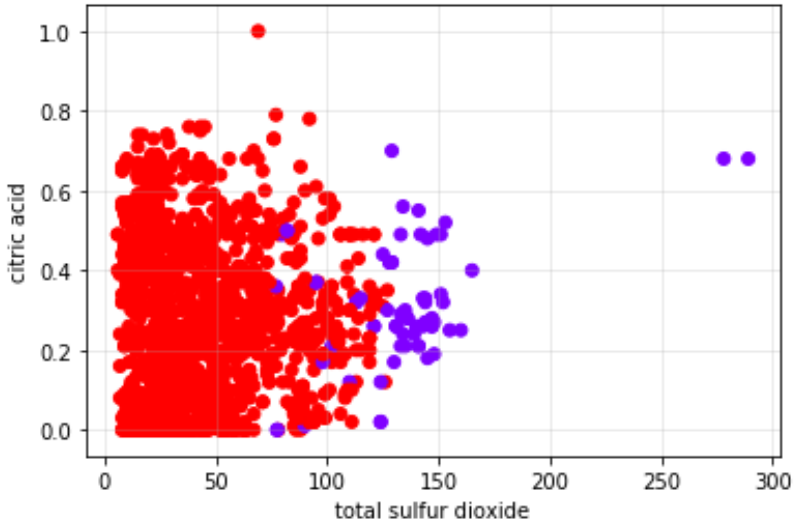
min-sample	Silhouette	Calinski-Harabasz	Davies-Bouldin
5	-0.4254	15.0433	1.6694
10	-0.6121	9.9393	2.3454
20	-0.4769	22.4975	1.0388
30	0.2597	112.3621	0.7562
40	0.5958	682.6572	0.5330

Tabela 24: Wartości metryk dla różnej liczby minimalnych próbek, dla zbioru B z wykorzystaniem algorytmu optics

min-cluster-size	Silhouette	Calinski-Harabasz	Davies-Bouldin
0.05	0.5958	682.6572	0.5330
0.1	0.5958	682.6572	0.5330
0.5	0.5958	682.6572	0.5330
2	0.5958	682.6572	0.5330
5	0.5958	682.6572	0.5330
10	0.5958	682.6572	0.5330
20	0.5958	682.6572	0.5330
30	0.5958	682.6572	0.5330

Tabela 25: Wartości metryk dla różnej liczby minimalnych próbek, dla zbioru B z wykorzystaniem algorytmu optics

Wartości metryk nie uległy zmianie pod wpływem dobierania różnych minimalnych wartości rozmiarów clastra. Metryki osiągnęły najlepsze wartości dla minimalnej liczby próbek wynoszącej 40



Rysunek 22: Wizualizacja wyliczonych centrów dla zbioru B przy pomocy optics

Zbiór C

Badania dla zbioru C były niemożliwe do przeprowadzenia dla niniejszej metody klasteryzacji, ponieważ program potrzebował zbyt dużej ilości pamięci.

6 Wnioski

Algorytm EM

- W związku z tym, że algorytm potrzebuje na wstępie informacji o liczbie centrów, wymagana jest wcześniejsza wiedza na temat zawartości zbioru oraz spodziewane klasy.
- Najlepsze konfiguracje różniły się między zbiorami danych, także ważnym elementem jest ich dobranie do każdego zbioru oddzielnie.
- Zwiększanie liczby iteracji powyżej pewnej wartości jest bezcelowe, ponieważ wymagana dokładność osiągana będzie zanim ustawiona liczba będzie osiągnięta.

Algorytm k-średnich

- W związku z tym, że algorytm potrzebuje na wstępie informacji o liczbie centrów, wymagana jest wcześniejsza wiedza na temat zawartości zbioru oraz spodziewane klasy.
- W przypadku wybranych zbiorów danych A, B oraz C zmiana parametrów takich jak liczba losowań centrów, maksymalne iteracje czy zmiana ziarna nie miała znacznego wpływu na uzyskane rezultaty klasyfikacji.
- Algorytm dąży do przypisania każdemu centrowi równej liczby elementów, co w określonych wypadkach może pogarszać wyniki.

Algorytm hierarchiczne aglomeracyjny

- W związku z tym, że algorytm potrzebuje na wstępie informacji o liczbie centrów, wymagana jest wcześniejsza wiedza na temat zawartości zbioru oraz spodziewane klasy.
- Dobrze radzi sobie przy zbiorze A, gdzie klastry są od siebie wyraźniej oddzielone niż w zbiorze B.
- Wrażliwy na wartości odstające.
- Nieoptymalny pod kątem obliczeniowym.

Metoda gęstościowa DBSCAN

- Algorytm przyzwoicie radzi sobie z klasteryzacją.
- Ze względu na specyfikacje algorytmu, aby upewnić się czy liczba klastrów jest zadowalająca, należy zastosować metryki w celu porównania wyników. Przekłada się to na różne wyniki (zależne od zastosowanej metryk) jak i na dodatkową pracę.

Algorytm optics

- Algorytm optics lepiej radzi sobie z dużymi zbiorami niż algorytm DBSCAN.
- Nie wymaga specjalnego dostrajania poszczególnych parametrów

Literatura

[1] *Prelekcja Christiana Henniga dotycząca oceny jakości klasteryzacji*, https://www.youtube.com/watch?v=Mf6MqIS2ql4

[2] *Popularne wartości ziarna dla algorytmów pakietu Scikit-learn*, https://scikit-learn.org/stable/glossary.html#term-random-state