Sprawozdanie: Techniki eksploracji danych wielowymiarowych

Jarmoc Maja, indeks: 274263

Cel ćwiczenia: przeprowadzenie analizy K - means z użyciem zbioru danych „Cereals" zawierającego 30 rodzajów płatków śniadaniowych wraz z informacjami o kaloriach, ilości białka, tłuszczu, błonnika, węglowodanów, cukrów (w gramach) i sodu (w miligramach); (porcja: 30g.) Dane zostały ówcześnie przygotowane: usunięcie danych odstających, transformacja (o ile tego wymagały).

Dane zostały poddane autoskalowaniu przy HCA oraz PCA (nie ponawiam tego kroku przy K - means).

Przy pomocy metody k-średnich zostanie utworzonych k różnych możliwie odmiennych skupień. Algorytm ten polega na przenoszeniu obiektów ze skupienia do skupienia tak długo aż zostaną zoptymalizowane zmienności wewnątrz skupień oraz pomiędzy skupieniami. Oczywistym jest, iż podobieństwo w skupieniu powinno być jak największe, zaś osobne skupienia powinny się maksymalnie od siebie różnić.

Pierwszym etapem algorytmu K – means jest ustalenie liczby skupień (k). Określiłam je za pomocą wykresu łokcia (Elbow method) - wykres k na sumę kwadratów odległości (Sum of Squared Errors). Dobrym modelem algorytmu K – means jest ten, który posiada stosunkowo niską wartość SSE i niską ilość klastrów (k).

$$\mathrm{SSE} = \sum_{\underbrace{i=1}_{\text{cluster}}}^{n} (\underbrace{y_i}_{\text{data point}} - \underbrace{\hat{y}_i}_{\text{centroid}})^2$$

SSE - Różnica między wartością zaobserwowaną a wartością przewidywaną.

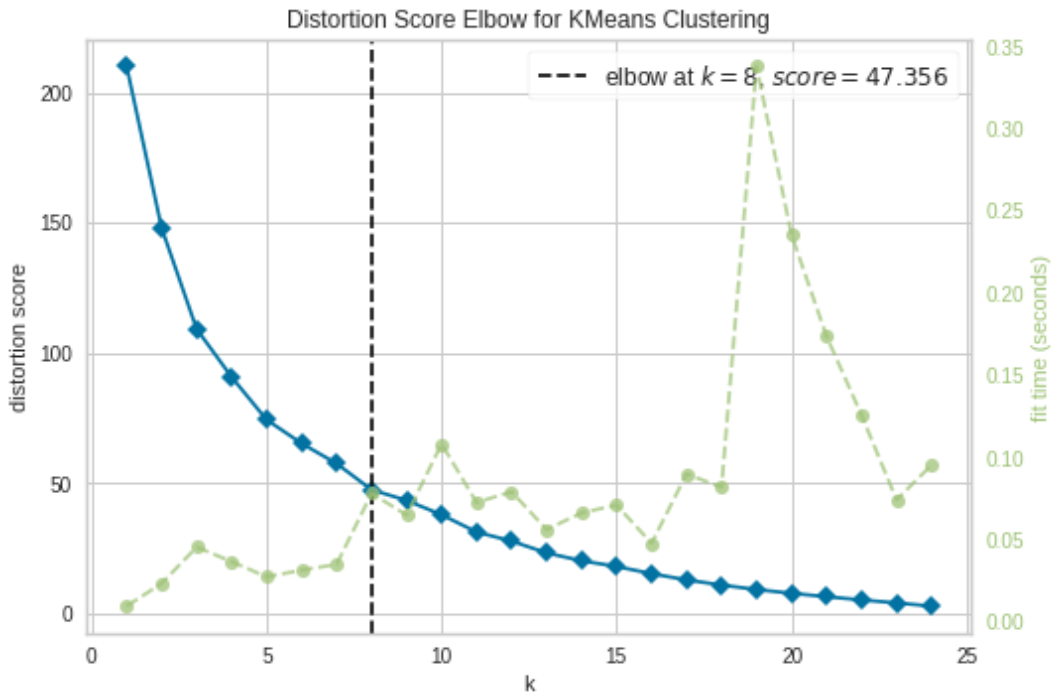


Figure 1. Elbow method with euclidean distance.

Odpowiednią wartością k przy moim zestawie danych jest k = 8 (dla odległości euklidesowej)

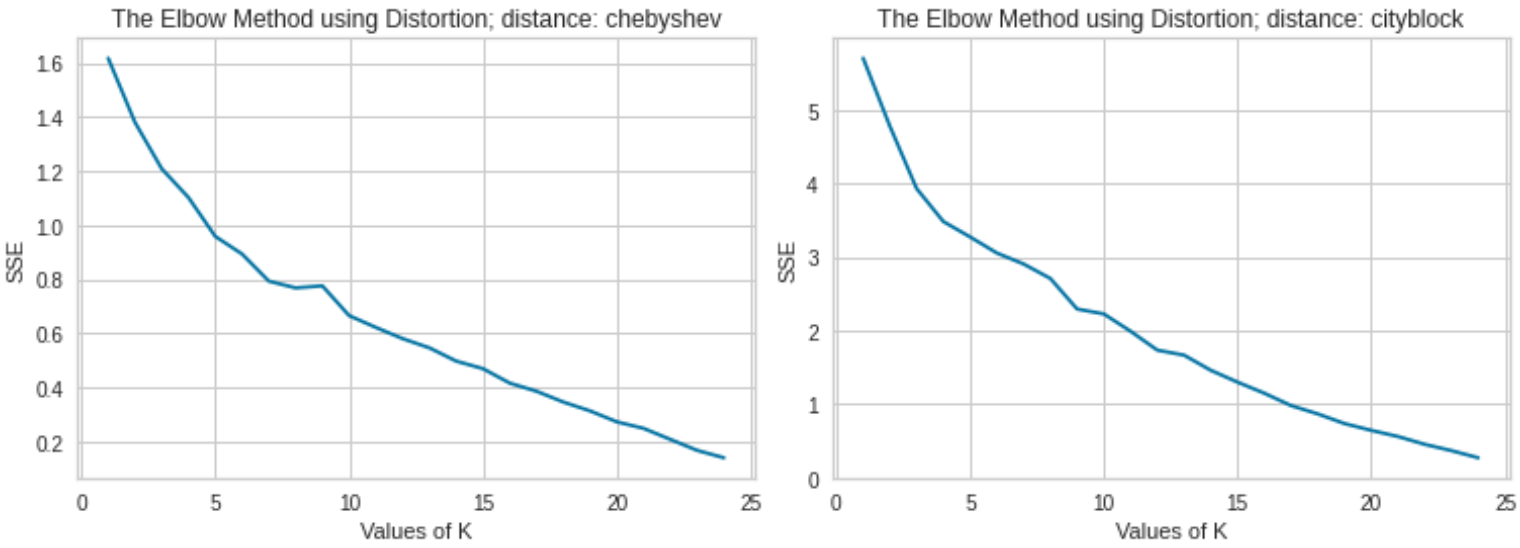


Figure 2. Elbow method with chebyshev and cityblock distance.

Następnie ustalamy wstępne środki skupień. Środki skupień, tak zwane centroidy, możemy dobrać na kilka sposobów: ja wybrałam losowo k centroidów.

Analizę przeprowadzam dla dwóch zmiennych – oś X – sód, oś Y – błonnik.

```
Centroid pierwszy: [0.8329, 1.4339]

Centroid drugi: [-0.5652, 0.2135]

Centroid trzeci: [-0.1956, -1.0070]

Centroid czwarty: [0.4882, -0.3966]

Centroid piąty: [-2.0382, 0.2135]

Centroid szósty: [-0.7519, -1.0068]

Centroid siódmy: [-2.0382, 0.82374]

Centroid ósmy: [0.1952, 2.0441]
```

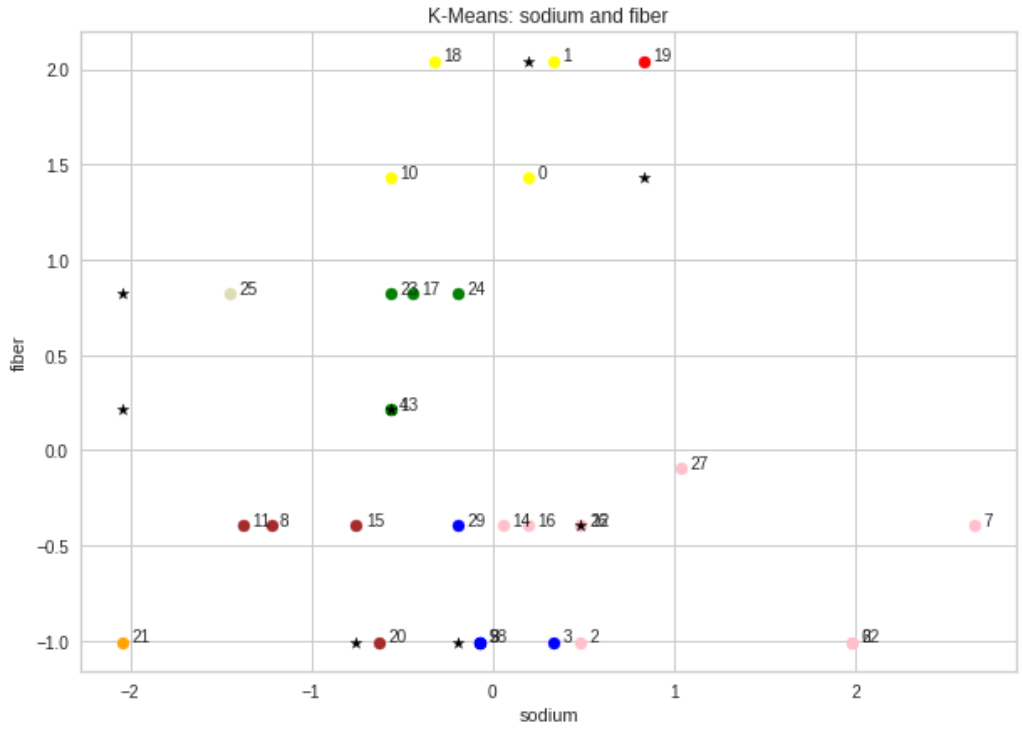


Figure 3. Położenie losowo wygenerowanych centroidów. Centroidy oznaczone gwiazdką.

Gdy mamy już współrzędne centroidów obliczamy odległość obiektów od środków skupień. W tym wypadku użyłam odległości euklidesowej.

Po dziesięciu iteracjach algorytmu położenie obiektów oraz centroidów wyglądało następująco:

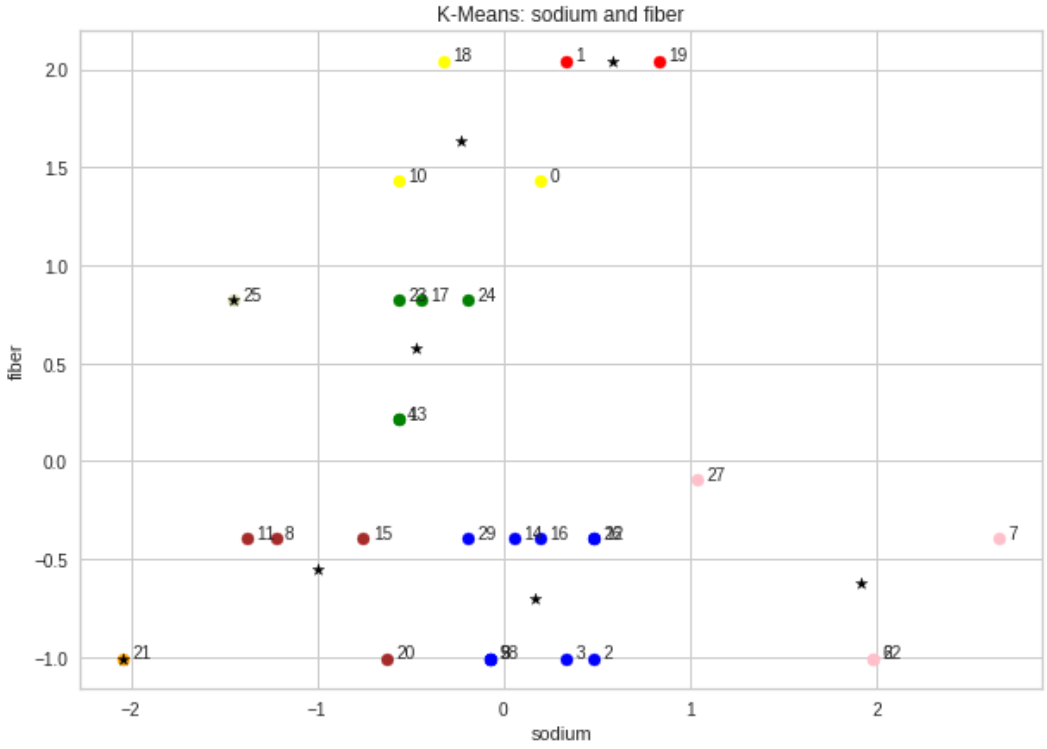


Figure 4. Położenie centroidów po 10 iteracjach algorytmu. Centroidy oznaczone gwiazdką.

Po trzecim przejściu algorytmu centroidy nie zmieniały swojego położenia.

Wyniki:

0	Bran Chex
1	Bran Flakes
2	Cap'n'Crunch
3	Cheerios
4	Cinnamon Toast Crunch
5	Clusters
6	Cocoa Puffs
7	Corn Chex
8	Corn Pops
9	Count Chocula
10	Cracklin' Oat Bran
11	Cream of Wheat (Quick)
12	Crispix
13	Crispy Wheat & Raisins
14	Double Chex
15	Froot Loops
16	Frosted Flakes
17	Frosted Mini-Wheats
18	Fruit & Fibre Dates; Walnuts; and Oats
19	Fruitful Bran
20	Fruity Pebbles
21	Golden Crisp
22	Golden Grahams
23	Grape Nuts Flakes
24	Grape-Nuts
25	Great Grains Pecan
26	Honey Graham Ohs
27	Honey Nut Cheerios
28	Honey-comb
29	Just Right Crunchy Nuggets

Płatki w klastrze czerwonym to Bran Flakes i Fruitful Bran – 5 gram błonnika oraz 210 i 240 mg sodu.

W klastrze żółtym znajdują się Bran Chex, Cracklin' Oat Bran, Fruit & Fibre Dates; Walnuts; and Oats. Płatki te mają 140 – 200 mg sodu oraz 4-5 gram błonnika. Skupisko różni się od wyżej wymienionego skupiska (czerwonego) rozbieżnością w zawartości sodu.

Płatki posiadające w nazwie słowo „bran” (otręby) charakteryzują się w większości przypadków dużą ilością błonnika w składzie. Cechą wyróżniającą otręby jest wysoka zawartość błonnika.

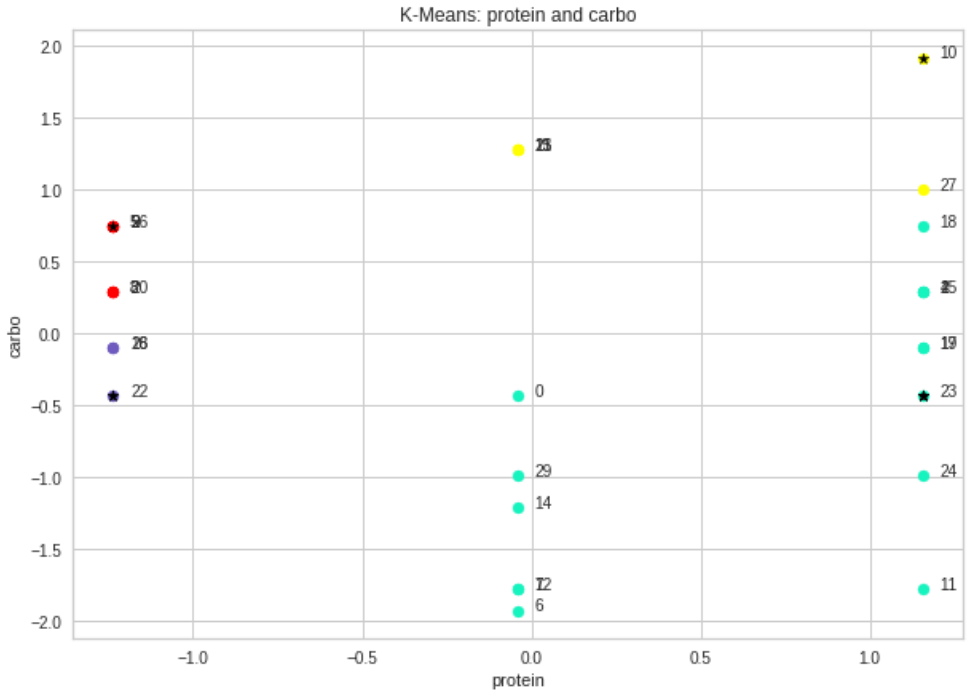
Płatki Great Grains Pecan (25) tworzą jednoelementowe skupisko (75 mg sodu i 3 gramy błonnika). Podobnie jak płatki Golden Crisp (21) – 45 mg sodu oraz brak błonnika.

Płatki w skupisku różowym – ponad 220 mg sodu.

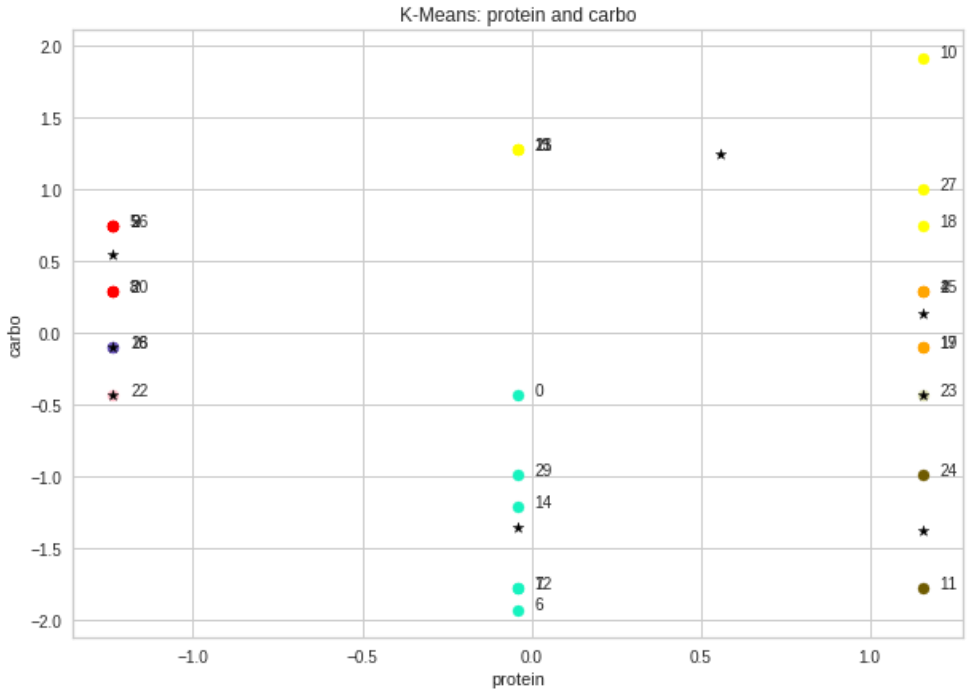
Suma błędów (SSE) dla współrzędnej x (sód) jest niższa niż dla współrzędnej y (błonnik).

Druga analiza dla x – białek i y – węglowodanów używając odległości Czebyszewa.

Przed:



Po:



Wyniki nie są już tak czytelne ze względu na nakładanie się na siebie punktów spowodowane identyczną ilością białka i węglowodanów w poszczególnych płatkach. Zbiór danych zawiera niedokładne informacje, często zaokrąglone do dziesiątek – stąd brak różnic. Po analizie mamy 2 klastry zawierające tylko jeden element (średnia z klastra równa jest elementowi)

Płatki 15 i 21 (Froot Loops i Golden Crisp) to płatki posiadające podobną ilość węglowodanów jak płatki znajdujące się z nimi w skupisku żółtym, różnią się jednak od nich ilością białka – pod względem tej cechy pasując bardziej do skupiska niebieskiego.

Suma błędów większa jest dla osi y (węglowodany) niż dla osi x (białka) – co jest widoczne także gołym okiem.

Wiekszą ilością białka charakteryzują się często płatki z słowem „nuts” / „walnuts” w nazwie (Grape-Nuts (24), Honey Nut Cheerios (27)). Orzechy oprócz dużej zawartości tłuszczu posiadają także sporą zawartość białka.