Podstawy Uczenia Maszynowego

Laboratorium 4

Maria Polak

1. **Przygotowanie zbioru**

Ze zbioru zostały usunięte następujące kolumny:

- Category (przeniesione do osobnej zmiennej dla późniejszej weryfikacji wyników)
- Item (nazwa nic nie wnosi do naszego przetwarzania)
- Calories (wartość wynika ze składu)
- Calories from Fat (wartość wynika z tłuszczy)
- Serving Size (nie jest to bardzo przydatna informacja oraz może prowadzić do redundancji, jeżeli np w menu pojawią sie takie same burgery w dwóch rozmiarach)
- Wszystkie kolumny typu “% Daily Value”, które mają odpowiedniki (kolumne nie daily)

Każda kolumna została znormalizowana za pomocą funkcji *preprocessing.normalize* z biblioteki *sklearn*. Zbiór nie został wycentrowany ponieważ nie jest to konieczne (w przeciwieństwie do PCA).

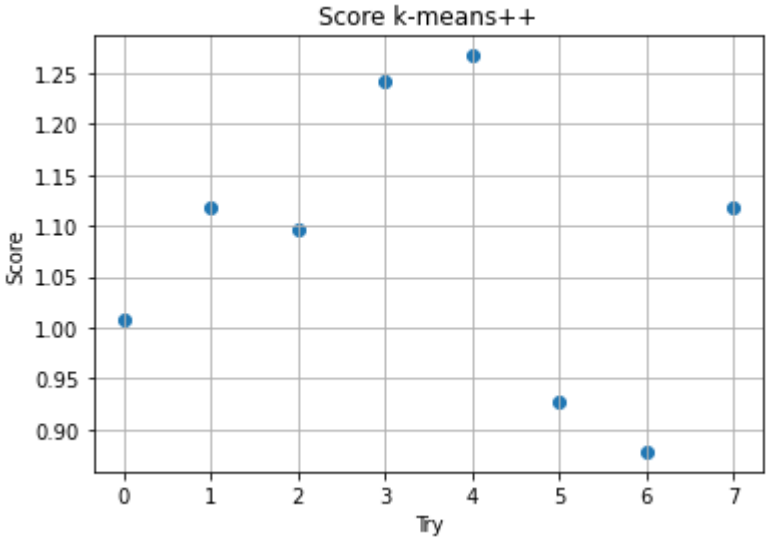
W wyniku tych przekształceń uzyskaliśmy dane o kształcie 260 wierszy x 13 kolumn z wartościami w zbliżonych przedziałach.

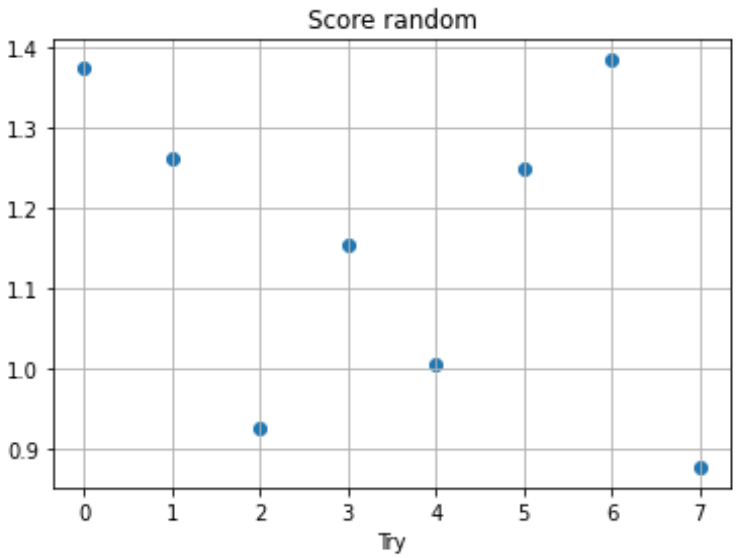
2. **Miara**

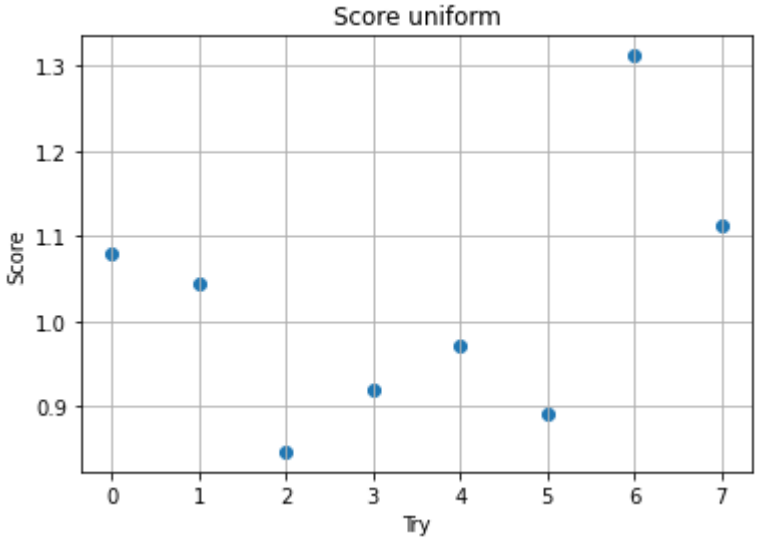
Na potrzeby zadania została wybrana miara Davies-Bouldina. Zwraca ona średnie podobieństwo każdego klastra do najbardziej mu podobnego. Podobieństwo jest definiowane jako stosunek dystansów w środku i między klastrami. Miare można używać tylko z metryką euklidesa (jest to związane z definicją średniej) i może ona zwrócić słabe wyniki w sytuacji, gdy klastry są nierówne względem promieni. Wynik im bliżej 0 = tym lepszy wynik.

3. Wpływ szczęścia

Dla każdgo sposobu inicjalizacji środków kmeans zostało wywołane 8 krotnie. Na poniższych wykresach widzimy, że wyniki bardzo się różnią między sobą i faktycznie zależą ‘od szczęścia’.

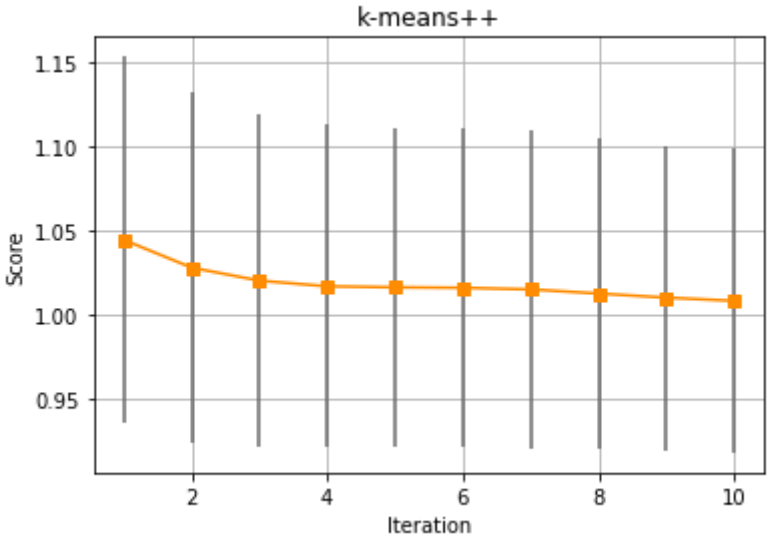


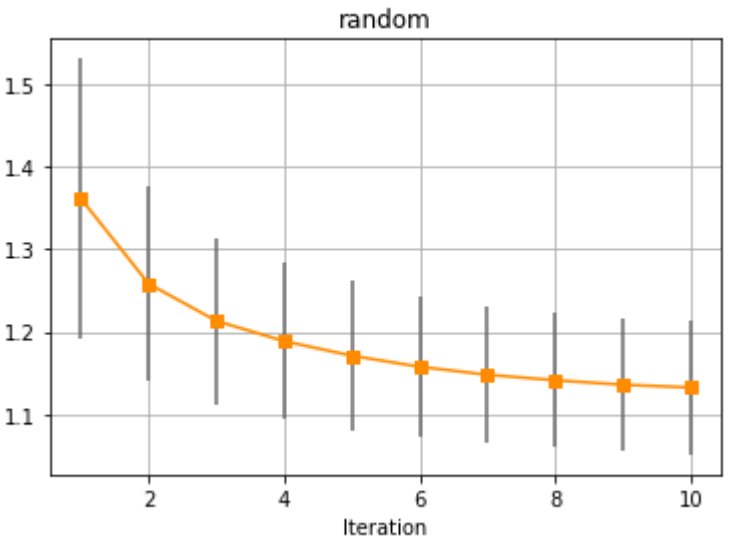


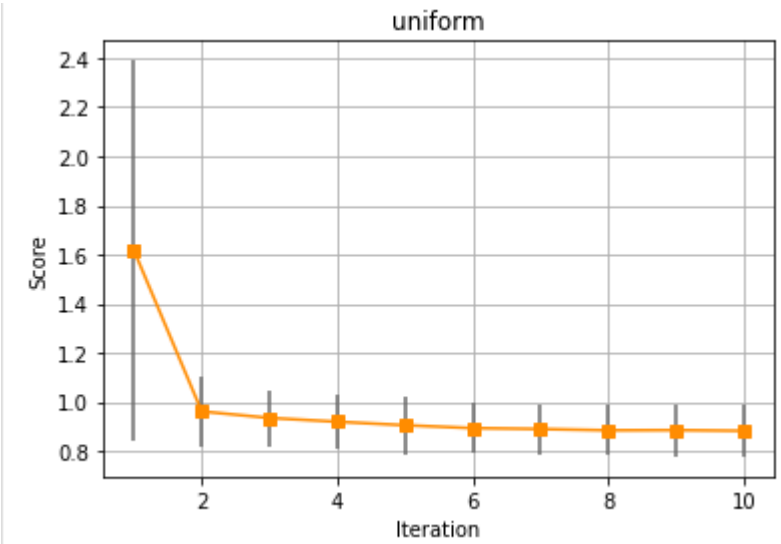


4. Proces Klasteryzacji

Proces klasteryzacji został zbadany krok po kroku. Testy zostały powtórzone 10 krotnie, aby wyeliminować element szczęścia.



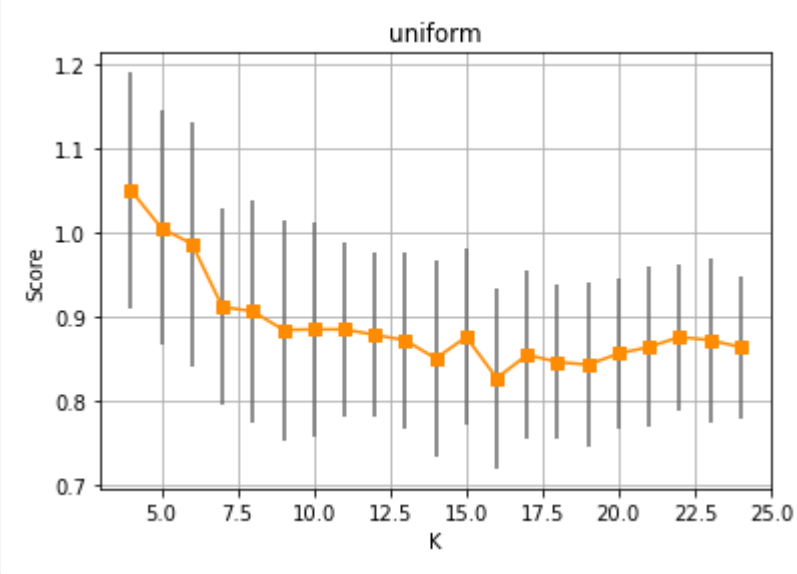




Losowanie środka z rozkładem jednostajnym zwróciło najlepsze wyniki.

5. **Ustalenie wartości K**

Dla wybranego losowania został przeprowadzony test, którego celem było wybranie najlepszego K.



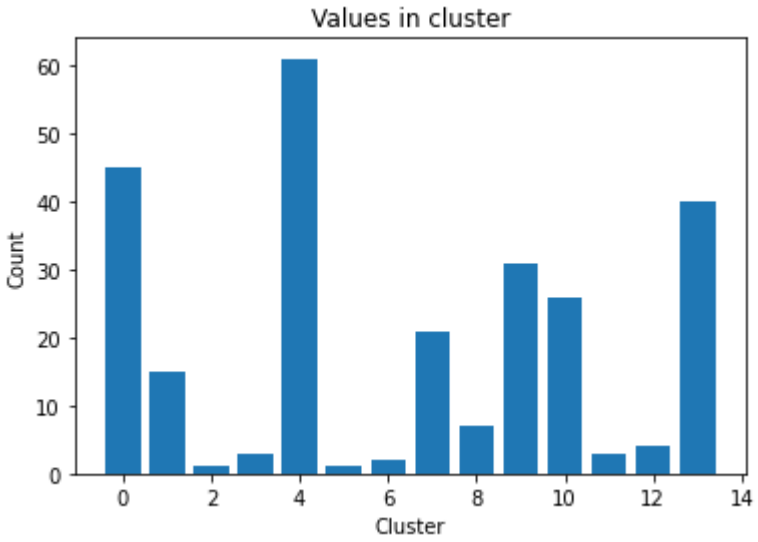
Jako najlepsze K została wybrana wartość 14. W tym miejscy score przestało opadać i zaczęło wyrównywać swoją wartość.

6. **Jakie klastry uzyskaliśmy?**

- Ile ich jest? Klastrów jest 14.
- Gdzie leżą ich środki? Bez wykresu ciężko powiedzieć ;)

```
array([[ 8.49221561e-03,  1.01397642e-02,  1.73472348e-17,
         4.01477142e-03,  4.67912361e-03,  1.66809277e-02,
         6.09801444e-03,  1.76258171e-02,  4.24073279e-03,
         5.40751395e-03,  4.63621803e-03,  5.00416476e-03,
         3.78684767e-03],
       [ 8.62286509e-02,  9.12578777e-02,  1.69967317e-01,
         8.19013370e-02,  9.30927965e-02,  4.63664768e-02,
         7.50055777e-02,  1.21869935e-02,  1.04840338e-01,
         2.06874616e-02,  1.67502071e-02,  5.74457689e-02,
         1.31356279e-01],
       [ 1.85659775e-01,  1.54674369e-01,  0.00000000e+00,
         3.46274035e-01,  1.84444524e-01,  1.30546391e-01,
         1.92087455e-01,  2.56833335e-02,  1.27221984e-01,
         3.34868983e-02,  4.48666261e-03,  6.89349227e-02,
         2.13010182e-01],
       [ 3.71319549e-02,  3.48017330e-02,  0.00000000e+00,
         3.61329428e-02,  5.03277831e-02,  1.50053323e-02,
         6.40291517e-02,  6.54673206e-03,  7.42128238e-02,
         2.82778253e-01,  6.35610536e-02,  3.44674614e-02,
         3.90518666e-02],
```

- Ile jest obserwacji w każdym klastrze?

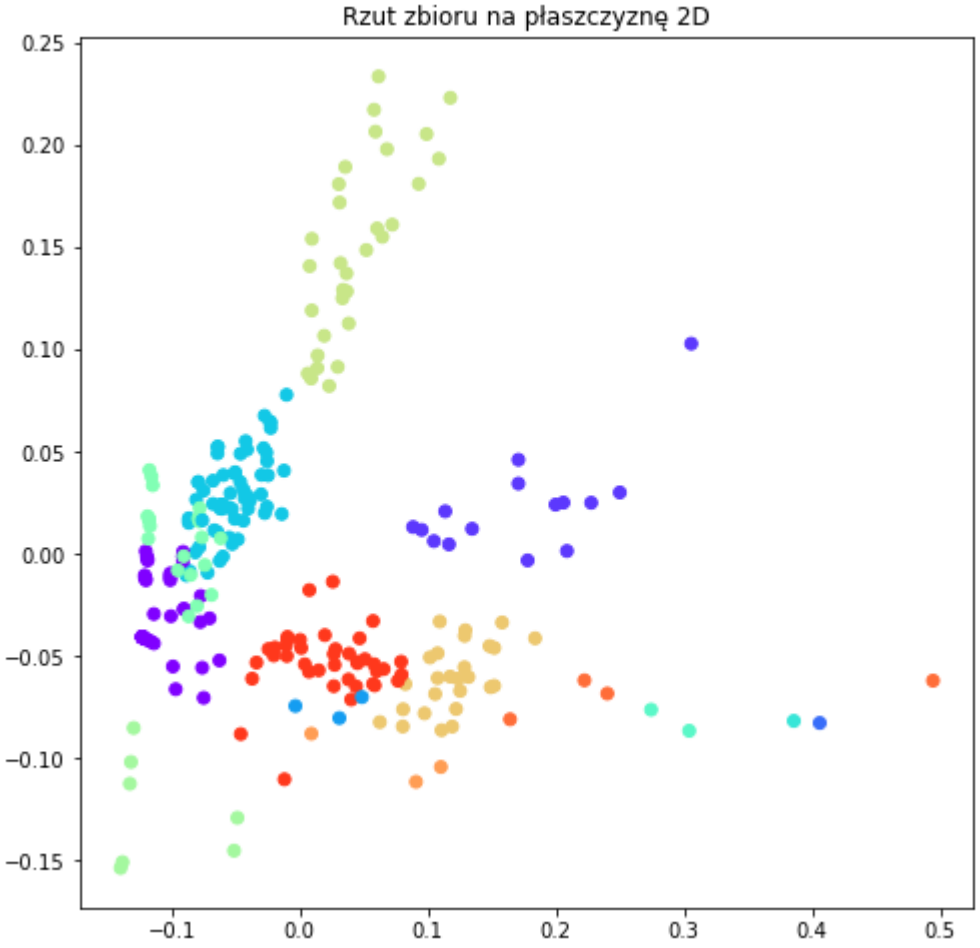


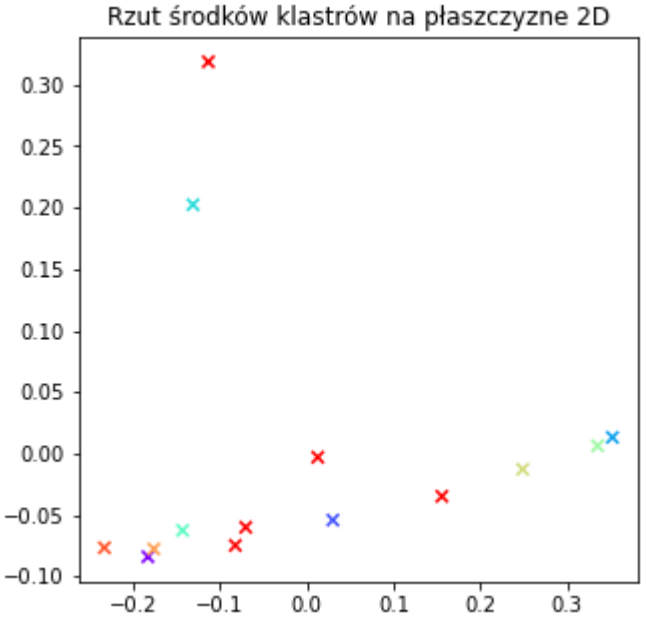
- Czy mają sens dla człowieka? Jak scharakteryzował(a)byś co znalazło się w każdym z nich?

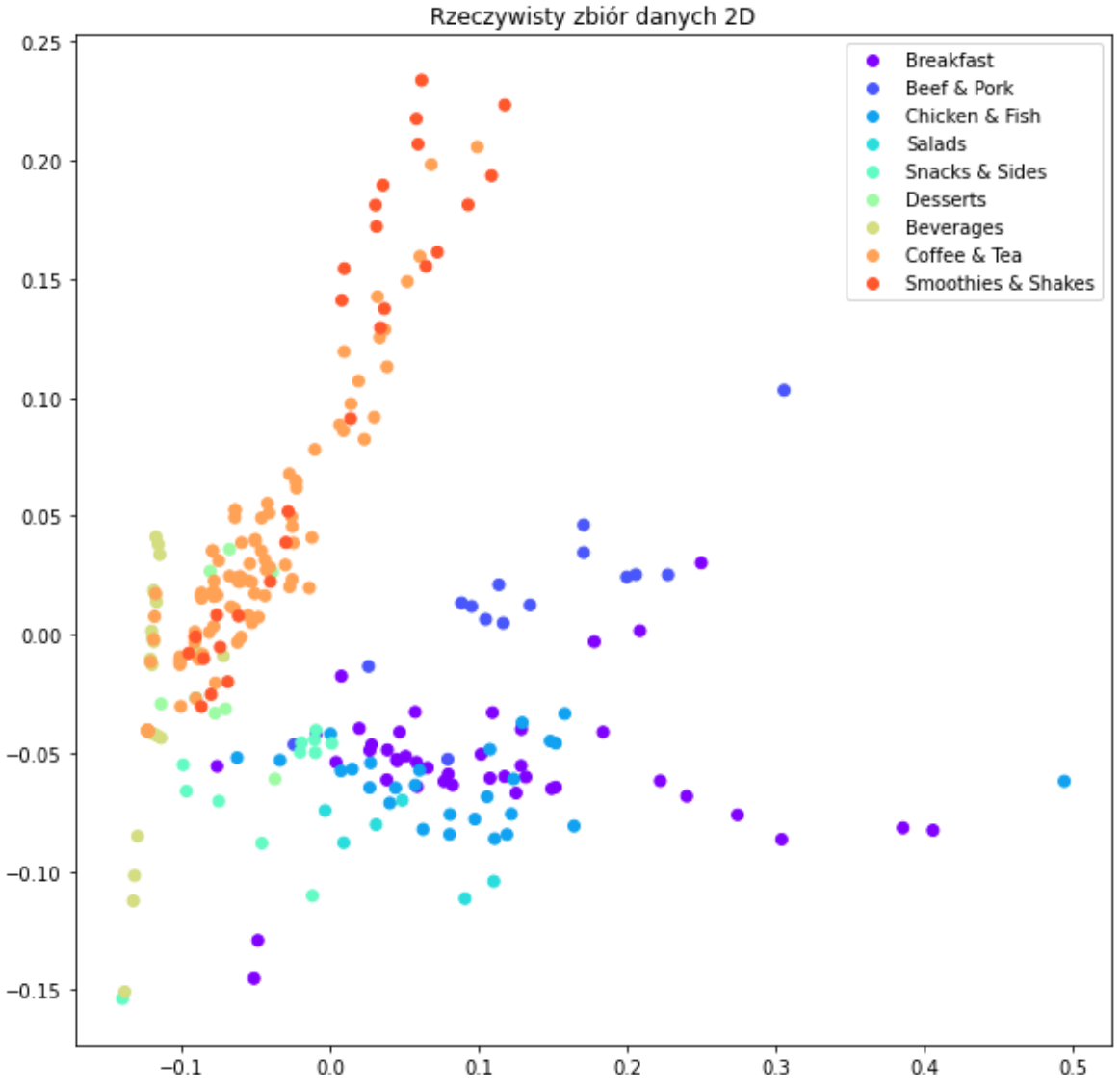
Jedzenie nie zostało podzielone na kateogrie podobne do ‘normalnych/ludzkich’ kategorii. Kmeans znajduje zależności między podanymi wartościami i na podstawie tego tworzy swoje kategorie - czasami trafne czasami nie.
Dwa przykłady z dużym uproszczeniem:

- Z perspektywy kmeans smoothie może być podobne do kawy z lodami ponieważ ma podobne ilości składników, mimo że dla nas są to dwie bardzo różne potrawy.
- Powiedzmy że w naszym menu główne dania można podzielić na dwie podkategorie burgery i makarony. Jesteśmy w stanie powiedzieć, że oba te typy dań są daniami głównymi, ponieważ tak jesteśmy nauczeni i ‘jemy na obiad’, jednak z perspektywy kmeans burgery i makarony drastycznie różnią się składem, więc on na pewno rozdzieli je na inne klastry.

7. Wizualizacja







Uzyskane klastry nie odpowiadają kategoriom z danych wejściowych, lecz niektóre fragmenty podziałów są zaskakująco zbliżone. W trakcie wykonywania zadania modyfikowałam klikukrotnie sposób przygotowania danych i wyniki przedstawione w ninjeszym raporcie (mimo że ogólnie nie są najlepsze) jestem w stanie określić jako dobre/satysfakcjonujące w porównaniu do pierwszych prób.