Podstawy Uczenia Maszynowego - Restauracja "Pod Złotymi Łukami"

Kamil Krzempek

1 Zbiór danych

1.1 Odrzucone kolumny

Z dostarczonego zbioru danych odrzucono następujące kolumny:

- **Category**, **Item** - ciężko byłoby określić mapowanie nazw na liczby, tak żeby oddawało relację między elementami. Przykładowo, nazwy *Cheeseburger* oraz *Jalapeño Double* odnoszą się do tego samego typu pożywienia, ale nie są do siebie podobne. Ponadto **Category** zostało wykorzystane do porównania uzyskanej klasteryzacji z rzeczywistymi kategoriami produktów
- **Serving Size** - różne jednostki, niektóre wartości odnoszą się do objętości, inne do masy
- **Calories from Fat**, **Total Fat**, **Saturated Fat**, **Cholesterol**, **Sodium**, **Carbohydrates**, **Dietary Fiber** - każda z tych kolumn ma odpowiednik określający % zalecanego dziennego spożycia. Wartości w tych kolumnach są wprost proporcjonalne, wartości procentowe wydają się jednak nieść bardziej użyteczną informację.

1.2 Skalowanie i normalizacja

Pozostałe kolumny zawierająco procentowe przeskalowano dzieląc wartości przez 100. **Calories** podzielono przez 2000, która to wartość jest interpretowana jako zalecane dzienne spożycie kalorii dla osoby dorosłej. Reszta z wartości została wycentrowana i znormalizowana tak, aby odchylenie standardowe było takie samo.

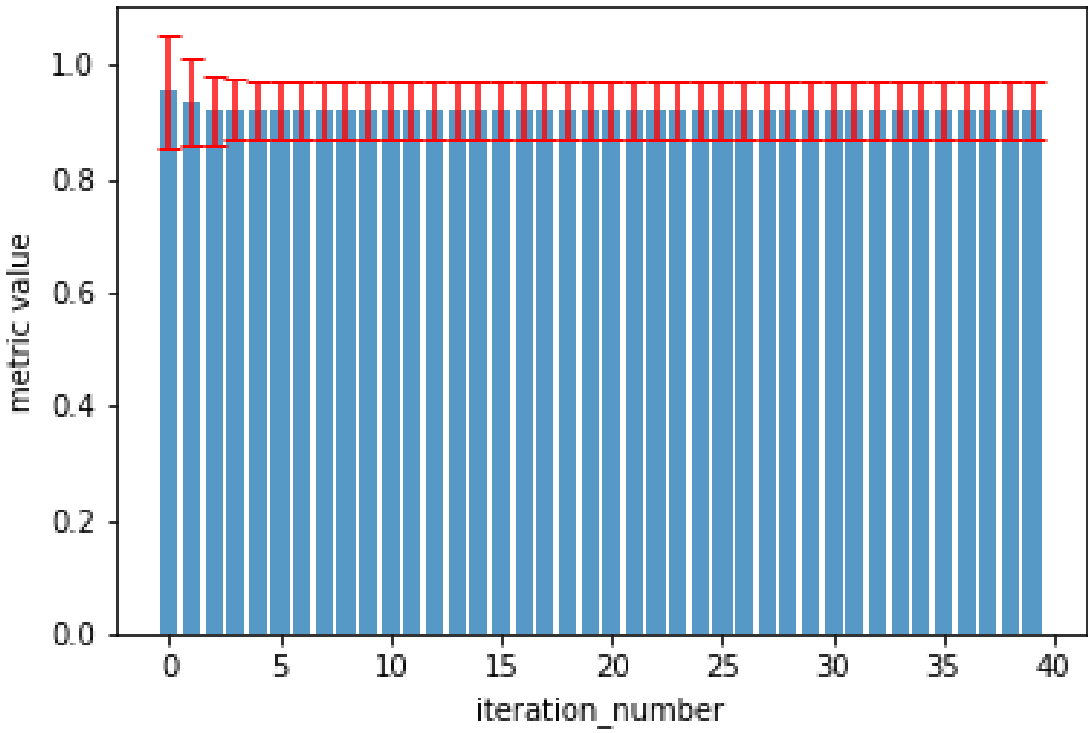
2 Wykorzystana metryka

Wykorzystano indeks Daviesa-Bouldina. Jest to metryka zdefiniowana jako średnie podobieństwo każdego z klastrów z klastrem najbardziej do niego podobnym. Podobieństwo jest rozumiane jako stosunek odległości wewnątrz klastra do odległości między klastrami. Klastry które są mniej rozproszone i bardziej oddalone od siebie otrzymają lepszy wynik w tej miarze.

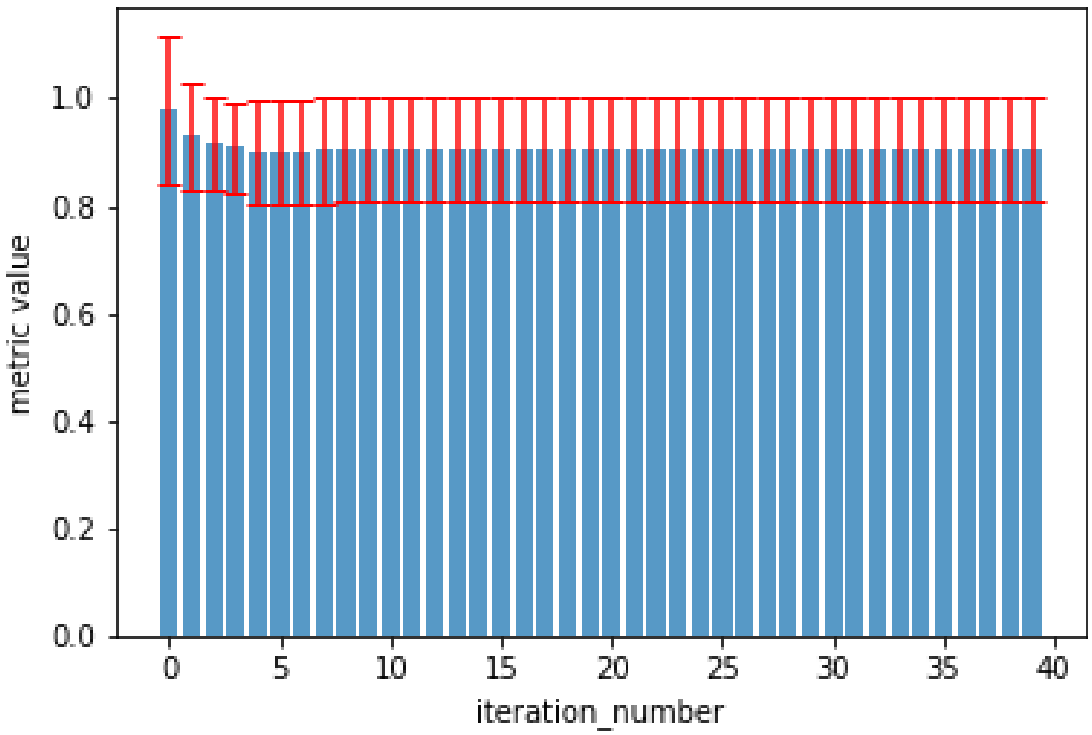
Ze względu na definicję tej miary, będzie ona faworyzować podziały z większą liczbą klastrów. Dla niektórych zbiorów danych bardziej sensowna może okazać się większa generalizacja, więc dla takich zbiorów nie byłaby to najlepsza miara.

3 Porównanie metod inicjalizacji klastrów

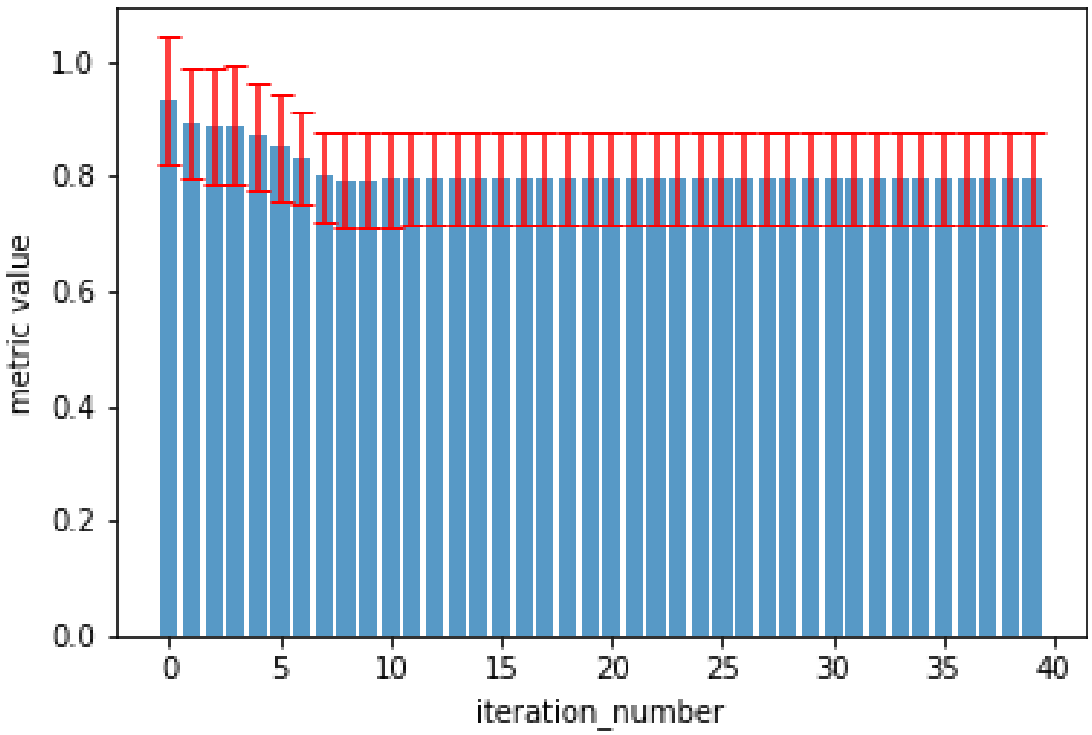
Przeprowadzono po 10 prób dla każdej metody, dla 6 klastrów i 40 iteracji.



Rysunek 1: Metoda k-means++, końcowa średnia = 0.920, końcowe odchylenie standardowe = 0.050



Rysunek 2: Metoda random, końcowa średnia = 0.908, końcowe odchylenie standardowe = 0.097

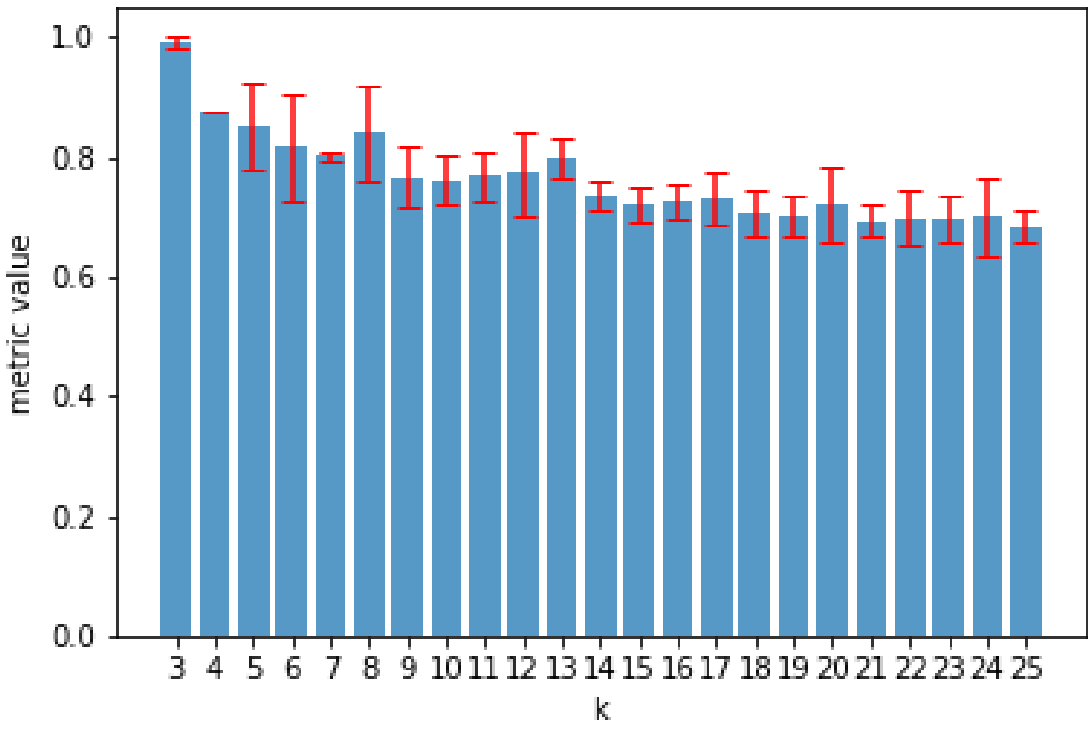


Rysunek 3: Własna implementacja wyboru środków, końcowa średnia = 0.798, końcowe odchylenie standardowe = 0.079

Własna implementacja wyboru środków dała najlepszy wynik.

4 Dobór K

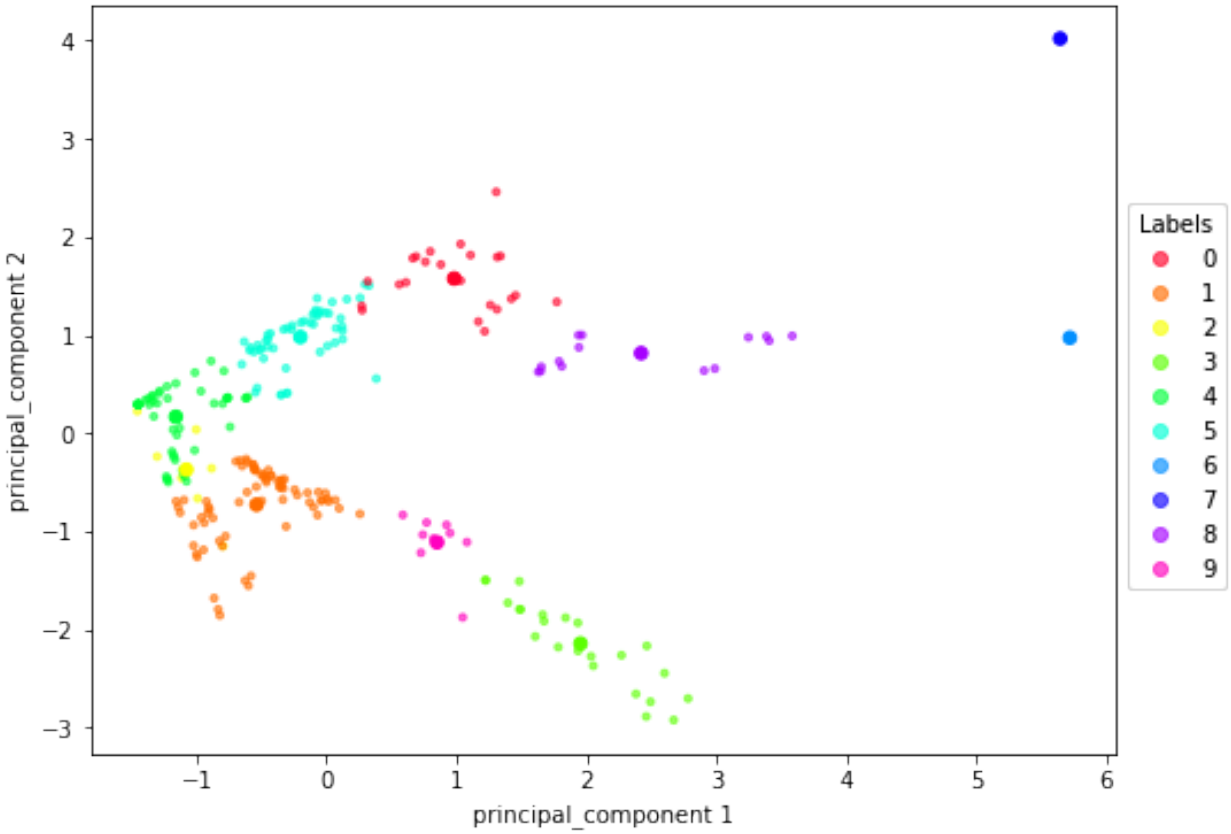
Dla każdej wartości K zostało przeprowadzone 10 prób.

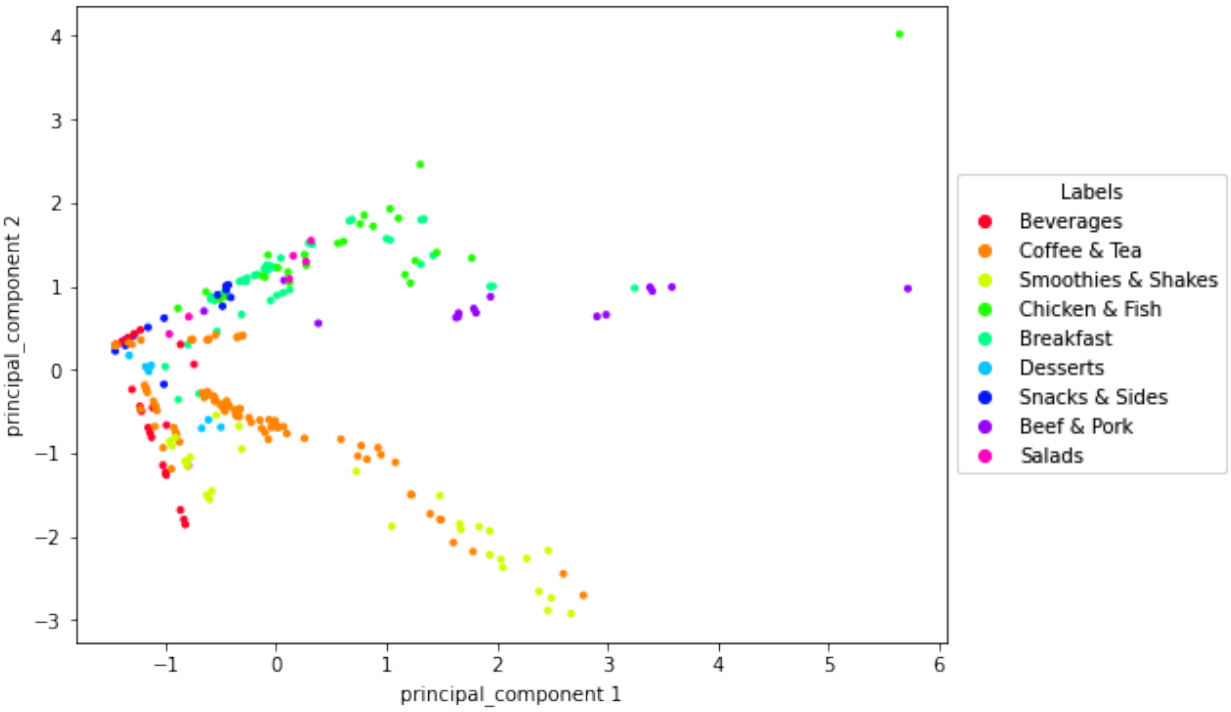


Zdecydowano się na wykorzystanie elbow rule i wybrano k = 10.

5 Wizualizacja klastrów

Dla k = 10, otrzymano 10 klastrów. Liczności poszczególnych klastrów bardzo się od siebie różniły, od 1 do 62. (liczności poszczególnych klastrów, w kolejności od najbardziej do najmniej licznego: 62, 55, 52, 36, 20, 17, 7, 6, 4, 1). Środki poszczególnych klastrów wydają się być dobrze dopasowane, mniej więcej na środku danego klastra. Występują pewne podobieństwa między otrzymaną klasteruzacją, a oryginalnymi kategoriami, ale ciężko je konkretnie zinterpretować.


principal_component 2
principal_component 1
Labels
0
1
2
3
4
5
6
7
8
9


principal_component 2
principal_component 1
Labels
Beverages
Coffee & Tea
Smoothies & Shakes
Chicken & Fish
Breakfast
Desserts
Snacks & Sides
Beef & Pork
Salads