Projekt - KMeans

Sebastian Michoń 136770, Marcin Zatorski 136834
grupe L5

1 Preprocessing

1. Kolumny 'name' i 'mfr' nie były używane w przetwarzaniu, kolumna 'type' została sprowadzona do zmiennej binarnej.

2. Atrybuty zostały znormalizowane (sprowadzone do przedziału <0;1> w standardowy sposób, formułą $\frac{x-min}{max-min}$)

3. Z pozostałych atrybutów nie były przetwarzane "weight", "shelf", "cups" i "ratings" - jeśli zadanie dotyczyło produktów podobnych ze względu na wartości odżywcze, to te atrybuty są prawdopodobnie zbędne.

2 Informacje wyróżniające powstałe grupy

1. W klastrach było kolejno: 14, 25 i 38 obserwacji

2. Obserwacje zawierające rodzynki (raisin w nazwie) były prawie wyłącznie w 3. grupie (9 z 11 obserwacji) podobnie jak musli (3 obserwacje) i miód (5 obserwacji). Kukurydza występowała głównie w 2. grupie (3 z 4 obserwacje), podobnie jak winogrona (2 obserwacje) i nasiona (4 z 5 obserwacji). W 1. grupie było 6 obserwacji z pszenicą w nazwie (były też 3 w 2. klastrze i 2 w 3. klastrze).

3. Obserwacje z 1. klastra posiadały średnio najwyższe oceny (minimum wyższe niż maksimum trzeciego klastra), zawierały najmniej witamin, sodu i tłuszczu.

4. W drugim klastrze obserwacje zawierały średnio najwięcej węglowodanów, sodu i witamin.

5. W trzecim klastrze obserwacje zawierały średnio najwięcej cukru, tłuszczu i kalorii (najniższa wartość kalorii była większa niż najwyższa wartość dla 1. grupy). Ponadto miały najniższe oceny.

6. Kilka składników odżywczych rozkładało się równomiernie między grupami - potas, białko, do pewnego stopnia błonnik.

3 Tezy dalsze

1. Zależnie od wybranych początkowych centroidów powstałe grupy są różne, przy czym pewne obserwacje zawsze występują razem.

2. Rezultaty własnego kMeansa były podobne do rezultatów tego samego algorytmu ze scikit-learn.