***** REPUBLIQUE DU BÉNIN *****
UNIVERSITE D'ABOMEY-CALAVI

ECOLE NATIONALE D'ÉCONOMIE APPLIQUÉE ET DE MANAGEMENT (ENEAM)

Page de garde

Filière : INFORMATIQUE DE GESTION (Analyse Informatique et Programmation)

Discipline:
INTRODUCTION A L'INTELLIGENCE ARTIFICIELLE

TP:
SEGMENTATION ~ MACHINE LEARNING ~ WEKA

Membres du Groupe 1

- BAWA SACCA Hamid
- COCOUVI Alexandro
- HOUNKPATIN Dèwanou Hugues-Marie
- OUSSA Chadrac Espoir
- PATINDE Nolan

Professeur
Dr DAGBA Théophile Komlan

ANNÉE ACADÉMIQUE

2025 – 2026

Sommaire

Table des matières

Introduction

Le présent projet s’inscrit dans le cadre de l’exploration de données appliquée à l'industrie agroalimentaire. L'objectif principal est de réaliser une segmentation automatique des produits céréaliers à l'aide de l'algorithme de partitionnement K-Means, en utilisant le jeu de données « Cereals » issu du dépôt UCI Machine Learning Repository.

La problématique centrale de ce travail consiste à identifier des structures cachées au sein du dataset en regroupant les céréales selon leurs similitudes nutritionnelles (calories, protéines, sucres, fibres) et commerciales (étagères). Pour ce faire, nous avons utilisé le logiciel WEKA afin de tester deux configurations distinctes : une segmentation fine en k=5 clusters et une approche plus globale en k=3. Cette étude nous permettra non seulement de caractériser chaque segment de marché, mais aussi de déterminer quel niveau de précision offre l'interprétation la plus pertinente pour un expert en nutrition.

I-Compréhension des données

Le jeu de données utilisé contient 77 instances dans le fichier (cereals.csv), chacune représentant une céréale, décrite par 16 attributs, dont 13 attributs numériques réellement exploités par l'algorithme.

I-1-Attributs utilisés dans l'analyse

Les attributs pris en compte par WEKA sont :

- Calories : apport énergétique
- Protéines (protein) : teneur en protéines
- Lipides (fat) : teneur en matières grasses
- Sodium : teneur en sel
- Fibres (fiber) : apport en fibres alimentaires
- Glucides (carbo) : quantité totale de glucides
- Sucres (sugars) : part des glucides simples
- Potassium (potass) : teneur en minéraux
- Vitamines (vitamins) : enrichissement en vitamines
- Durée de conservation (shelf) : position sur l'étagère (indicateur commercial)
- Poids (weight) : masse d'une portion
- Contenance (cups) : volume d'une portion
- Note (rating) : appréciation globale des céréales

I-2-Attributs ignorés

Les attributs suivants ont été exclus car non numériques et donc non pertinents pour le clustering :

- name
- mfr
- type

Ainsi, la segmentation repose uniquement sur des critères mesurables et comparables, ce qui garantit la cohérence des résultats.

II-Prétraitement des données

Avant l'application de l'algorithme de clustering, deux étapes essentielles de prétraitement ont été réalisées dans WEKA afin d'assurer la qualité de l'analyse.

III-1 Gestion des valeurs manquantes : ReplaceMissingValues

Le filtre ReplaceMissingValues a été appliqué pour traiter les champs vides dans le dataset.

- Pour les attributs numériques, WEKA remplace les valeurs manquantes par la moyenne de l'attribut
- Cette étape évite :
    - la perte d'instances,
    - les erreurs de calcul,
    - et les biais dans la formation des clusters

Ainsi, aucune donnée n'est supprimée et l'ensemble des 77 céréales est conservé.

III-2 Normalisation des données : Normalize

Le filtre Normalize a ensuite été appliqué.

- Toutes les valeurs numériques sont ramenées dans l'intervalle [0 ; 1]
- Cela permet de :
    - rendre comparables des attributs de natures différentes (ex : calories vs vitamines),
    - éviter qu'un attribut à grande échelle domine les autres,
    - améliorer la stabilité et la pertinence du clustering

IV-Implémentation dans WEKA

IV-1-Algorithme k-mean avec k=5

Paramétrage sous WEKA

- Algorithme : SimpleKMeans
- Nombre de clusters : k = 5
- Distance : Euclidienne
- Prétraitement :
    - ReplaceMissingValues : remplacement des valeurs manquantes par la moyenne (numérique)
    - Normalize (0–1) : mise à l'échelle homogène de tous les attributs
- Attributs utilisés :

    - calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight, cups, rating
- Attributs ignorés : name, mfr, type
- Nombre d'instances : 77

Résultats pour k = 5

Voici les images des résultats présentés par WEKA, nous pouvons également consulter le fichier: cereals_results_k5.txt

18:53:20 - SimpleKMeans

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     cereals-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    77
Attributes:   16
              calories
              protein
              fat
              sodium
              fiber
              carbo
              sugars
              potass
              vitamins
              shelf
              weight
              cups
              rating
Ignored:
              name
              mfr
              type
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 25.71312405440395

Initial starting points (random):

Cluster 0: 0.454545,0.2,0,0.90625,0.071429,0.916667,0.1875,0.108761,0.25,0,0.5,0.6,45.863324
Cluster 1: 0.181818,0.6,0.2,0.40625,0.714286,0.25,0.4375,0.848943,0.25,1,0.5,0.064,68.402973
Cluster 2: 0.545455,0,0.2,0.5625,0,0.541667,0.875,0.199396,0.25,0.5,0.5,0.6,22.396513
Cluster 3: 0.636364,0,0.6,0.65625,0,0.583333,0.625,0.138973,0.25,0.5,0.5,0.4,19.823573
Cluster 4: 0.636364,0.4,0.2,0.625,0.428571,0.5,0.9375,0.78852,0.25,1,0.83,0.336,37.840594
```

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                        Cluster#
Attribute    Full Data          0          1          2          3          4
                (77.0)     (19.0)      (7.0)     (20.0)     (17.0)     (14.0)
==============================================================================
calories        0.5171     0.4545     0.1558     0.5591     0.4973     0.7468
protein         0.3091     0.3895     0.3714        0.1     0.3412     0.4286
fat             0.2026     0.1053     0.0571        0.2     0.2118        0.4
sodium           0.499     0.4646     0.3304     0.5313     0.5533     0.5179
fiber           0.1537     0.1342     0.4184     0.0357     0.1282     0.2474
carbo           0.6499     0.7566     0.4643     0.5625     0.7071     0.6533
sugars          0.4951     0.2237     0.2589     0.7844      0.375     0.7143
potass          0.2933     0.2438     0.5619     0.1352     0.2436     0.5123
vitamins        0.2825     0.1842     0.1786       0.25     0.4265     0.3393
shelf           0.6039     0.1316          1       0.35     0.9706     0.9643
weight          0.5296     0.4911     0.3571        0.5        0.5     0.7464
cups            0.4568     0.5592     0.2949     0.5052     0.3944     0.4057
rating         42.6657    53.0023    64.8489    28.2854    44.1333    36.3072




Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      19 ( 25%)
1       7 (  9%)
2      20 ( 26%)
3      17 ( 22%)
4      14 ( 18%)
```

Répartition des clusters :

- Cluster 0 : 19 instances (25%)
- Cluster 1 : 7 instances (9%)
- Cluster 2 : 20 instances (26%)
- Cluster 3 : 17 instances (22%)
- Cluster 4 : 14 instances (18%)

Analyse détaillée des clusters (k = 5)

Cluster 0 (25%)

- calories = 0.4545
- protein = 0.3895
- fat = 0.1053
- sodium = 0.4646
- fiber = 0.1342
- carbo = 0.7566
- sugars = 0.2237
- potass = 0.2438
- vitamins = 0.1842
- shelf = 0.1316
- weight = 0.4911
- cups = 0.5592
- rating = 53.0023

Interprétation

Ce cluster représente 25 % de L'échantillon, ce qui en fait l'un des plus gros segments du marché. Avec un Rating de 53.0023, il offre une qualité nutritionnelle solide et supérieure à la moyenne.

1. Des céréales axés sur les Glucides (Carbo)

- Carbohydrates / Carbo (0.7566) : C'est la valeur la plus élevée de tous tes clusters. Ces céréales sont composées presque exclusivement de sucres complexes (glucose), fournissant une énergie de longue durée.
- Calories (0.4545) : Le niveau calorique est modéré, bien en dessous du Cluster 4 (le plus lourd).
- Sugars (0.2237) : C'est un excellent point : le taux de sucre est très bas. C'est ce qui explique son très bon Rating.

2. Composition Équilibrée

- Protéines (0.3895) : Un bon apport protéique, comparable aux céréales spécialisées.
- Fat (0.1053) : Très peu de matières grasses, ce qui renforce l'aspect sain du produit.
- Vitamines (0.1842) : Faible taux de vitamines ajoutées, indiquant souvent des produits simples (type riz soufflé ou flocons de maïs nature) peu transformés industriellement.

3. Positionnement Marketing

- Shelf (0.1316) : Ces céréales sont placées sur l'étagère du bas (étagère 1). Ce sont souvent des produits de basse étagère.
- Cups (0.5592) : Ce sont les céréales les plus légères en volume. Une portion remplit bien le bol sans être trop lourde en poids.

Conclusion: Le Cluster 0 regroupe les céréales dites “énergétiques”. Son profil se caractérise par un apport massif en glucides (le plus élevé du dataset) couplé à un taux de sucre très faible. C'est cette simplicité nutritionnelle qui lui permet d'obtenir un excellent score d'évaluation (Rating de 53.0023), le classant au deuxième rang des céréales les plus saines. Pour un consommateur, c'est le choix idéal pour un apport en énergie durable sans les effets négatifs du sucre ajouté.

Cluster 1 (9%)

- calories = 0.1558
- protein = 0.3714
- fat = 0.0571
- sodium = 0.3304
- fiber = 0.4184
- carbo = 0.4643
- sugars = 0.2589
- potass = 0.5619
- vitamins = 0.1786
- shelf = 1
- weight = 0.3571
- cups = 0.2949
- rating = 64.8489

Interprétation
Ce cluster représente 9 % du dataset, ce qui correspond à une niche de produits très spécifiques. Avec un Rating de 64.85, il s'agit du groupe de céréales ayant la meilleure qualité nutritionnelle globale.

1. Analyse des Macronutriments (Énergie et Construction du corps)

- Calories (0.1558) : C'est l'un des points les plus bas du cluster. Ces céréales sont très peu caloriques, ce qui influence positivement le rating.
- Protéines (0.3714) : Un taux modéré à élevé. Elles offrent une base protéique solide pour qu'il faut en tant que céréales.
- Lipides / Fat (0.0571) : La valeur est extrêmement faible. Ce sont des céréales quasiment sans matières grasses.

2. Profil Santé et Transit (Le point fort de ce cluster)

- Fibres (0.4184) : Ce score est très élevé par rapport à la moyenne. Ces céréales sont riches en fibres, favorisant la satiété et le transit.
- Potassium / Potass (0.5619) : C'est la valeur la plus haute de ce profil. Le potassium est souvent lié à la présence de sons ou de fruits secs, renforçant l'aspect "santé".
- Sucres (0.2589) : Le taux de sucre est maîtrisé (dans le quartile inférieur), ce qui est crucial pour maintenir un rating élevé.

3. Micro-nutriments et Minéraux

- Sodium (0.3304) : Le sel est présent de manière modérée, sans être excessif.
- Vitamines (0.1786) : Étonnamment, le score en vitamines ajoutées est bas. Cela suggère que ces céréales tirent leur valeur nutritionnelle de leurs ingrédients naturels plutôt que d'un enrichissement artificiel.

4. Logistique et Packaging

- Shelf (1) : Ces céréales se trouvent sur le premier rayon (souvent le bas). Ce n'est pas la position marketing la plus visible, ce qui confirme un ciblage "adulte/santé".
- Poids et Volume (Weight 0.3571 / Cups 0.2949) : Le rapport indique des céréales denses. Une petite quantité en volume (cups) pèse lourd.

Conclusion : Le Cluster 1 regroupe les céréales dites “diététiques “. Son profil est caractérisé par un équilibre optimal : très peu de calories et de graisses, mais un fort apport en fibres et en potassium. C'est cette combinaison spécifique qui lui permet d'atteindre le score d'évaluation (Rating) le plus élevé du dataset (64.85). Pour un consommateur, c'est le choix le plus sain identifié par l'algorithme K-Means.

Cluster 2 (26%)

- calories = 0.5591
- protein = 0.1
- fat = 0.2
- sodium = 0.5313
- fiber = 0.0357

- carbo = 0.5625
- sugars = 0.7844
- potass = 0.1352
- vitamins = 0.25
- shelf = 0.35
- weight = 0.5
- cups = 0.5052
- rating = 28.2854

Interprétation
Ce cluster est beaucoup plus important en volume (26 % du dataset). Avec un Rating très bas de 28.29, il regroupe les céréales les moins saines, souvent destinées aux enfants ou positionnées comme des gourmandises.

1. Analyse des Macronutriments (Énergie vide)

- Calories (0.5591) : Le taux est élevé (plus du triple du Cluster 1). Ces céréales sont denses en énergie mais pauvres en nutriments essentiels.
- Sucres (0.7844) : C'est la donnée la plus frappante. Ce cluster contient les céréales les plus sucrées du marché. C'est le facteur principal qui fait chuter la note de qualité (rating).
- Protéines (0.1) : Le taux est très bas, montrant que ces céréales n'offrent pas une base de construction solide pour l'organisme.

2. Profil Nutritionnel (Carences en éléments naturels)

- Fibres (0.0357) : Quasiment inexistantes. Ce sont des céréales hautement transformées, souvent à base de farine blanche raffinée.
- Potassium (0.1352) : Très faible, confirmant l'absence de grains entiers ou de fruits.
- Sodium (0.5313) : Un taux de sel assez élevé, souvent utilisé comme exhausteur de goût pour compenser la pauvreté des ingrédients de base.

3. Caractéristiques de consommation et Marketing

- Shelf (0.35) : Ce score indique que ces céréales sont souvent placées sur les étagères du milieu (étagère 2). C'est stratégique : c'est le niveau des yeux des enfants pour favoriser l'achat impulsif.
- Vitamines (0.25) : Le score est plus élevé que dans le Cluster 1. Cela indique souvent un enrichissement artificiel (vitamines ajoutées industriellement) pour essayer de compenser le manque de nutriments naturels.
- Cups (0.5052) : Ces céréales sont légères et volumineuses (type soufflées). On remplit davantage le bol, mais on consomme principalement du sucre et de l'air.

Conclusion: Le Cluster 2 (26 % des céréales) représente le segment “Moins sains”. Son profil est marqué par un déséquilibre nutritionnel majeur : un taux de sucre et de calories très élevé associé à une absence quasi totale de fibres et de protéines. Ce manque de nutriments essentiels explique son score de qualité extrêmement faible (Rating de 28.29). L'algorithme a ici parfaitement isolé les produits les moins recommandables d'un point de vue diététique.

Cluster 3 (22%)

- calories = 0.4973
- protein = 0.3412
- fat = 0.2118
- sodium = 0.5533
- fiber = 0.1282
- carbo = 0.7071
- sugars = 0.375
- potass = 0.2436
- vitamins = 0.4265
- shelf = 0.9706
- weight = 0.5
- cups = 0.3944
- rating = 44.1333

Interprétation:

Ce cluster représente 22 % du dataset. Avec un Rating de 44.13, il définit la "moyenne" du marché : des céréales classiques, souvent enrichies, qui ne sont ni excellentes ni mauvaises pour la santé.

1. Apport Énergétique et Glucides

- Calories (0.4973) : Une valeur moyenne-haute, assez proche du Cluster 2.
- Carbohydrates / Carbo (0.7071) : C'est la valeur la plus élevée de tous les clusters jusqu'ici. Cela indique des céréales très riches en amidon (maïs, blé soufflé), fournissant de l'énergie rapide.
- Sugars (0.375) : Le taux de sucre est modéré. C'est beaucoup moins que le Cluster 2 (0.78), ce qui explique pourquoi la note (rating) remonte à 44.

2. Équilibre Nutritionnel

- Protéines (0.3412) : Un bon niveau de protéines, presque aussi élevé que le cluster santé.
- Vitamines (0.4265) : C'est le point fort de ce groupe. Ce sont des céréales massivement enrichies en vitamines et minéraux pour compenser un profil de base moyen.

- Sodium (0.5533) : Le taux de sel est le plus élevé des trois clusters. Le sodium est utilisé ici pour donner du goût à des céréales riches en glucides complexes.

3. Positionnement Marketing

- Shelf (0.9706) : Proche de 1, ce qui signifie que ces céréales sont placées sur l'étagère du haut (étagère 3). C'est l'étagère destinée aux adultes qui cherchent des produits classiques ou des marques nationales reconnues (type Corn Flakes ou Special K).
- Fiber (0.1282) et Potass (0.2436) : Des valeurs modestes. Elles contiennent un peu de fibres mais restent loin des produits "grains entiers".

Conclusion:

Elles se distinguent par un taux de glucides très élevé (0.70) et un fort enrichissement en vitamines (0.42). Bien qu'elles soient assez salées, leur teneur en sucre reste raisonnable, ce qui leur permet d'obtenir un Rating correct de 44.13. C'est le segment de consommation courante, privilégiant l'énergie et les apports vitaminiques artificiels plutôt que les fibres naturelles.

Cluster 4 (18%)

- calories = 0.7468
- protein = 0.4286
- fat = 0.4
- sodium = 0.5179
- fiber = 0.2474
- carbo = 0.6533
- sugars = 0.7143
- potass = 0.5123
- vitamins = 0.3393
- shelf = 0.9643
- weight = 0.7464
- cups = 0.4057
- rating = 36.3072

Interprétation:

Ce cluster représente 18 % de ton échantillon. Avec un Rating de 36.31, il se situe dans la partie basse du classement, juste au-dessus du cluster des enfants (Cluster 2), mais pour des raisons différentes.

1. Un profil "Lourd" (Densité Énergétique)

- Calories (0.7468) : C'est le score le plus élevé de tous les clusters. Ces céréales sont très énergétiques.

- Weight (0.7464) : Le poids par portion est également le plus élevé. Ce sont des céréales lourdes et denses.
- Fat (0.4) : Le taux de matières grasses est le plus important ici. Cela confirme la présence d'ingrédients gras.

  2. Le paradoxe nutritionnel (Richesse et Excès)

- Protéines (0.4286) et Potassium (0.5123) : Les taux sont excellents. On y trouve beaucoup de nutriments essentiels.
- Sugars (0.7143) : C'est le point faible. Malgré les bons nutriments (protéines/fibres), le taux de sucre est presque aussi élevé que dans le cluster 2.
- Fiber (0.2474) : Un taux correct de fibres, mais qui ne suffit pas à compenser l'apport massif de sucre et de calories.

  3. Positionnement Marketing

- Shelf (0.9643) : Elles sont placées sur l'étagère du haut (étagère 3). Ce sont des produits "adulte" qui justifient leur positionnement élevé sur étagère.
- Rating (36.31) : Le score est bas car l'algorithme de notation pénalise lourdement le mélange Sucre + Gras + Calories, même si les protéines sont présentes.

  Conclusion:

  Le Cluster 4 (18 %) correspond au segment 'Énergie'. Ce groupe se distingue par les valeurs de calories, de gras et de poids les plus élevées du dataset. Bien qu'elles soient riches en protéines et en potassium, leur teneur en sucre très importante (0.71) fait chuter leur Rating à 36.31. Ce sont typiquement des céréales enrichis, perçus comme 'sains' par le consommateur mais très chargés d'un point de vue nutritionnel.

IV-2-Algorithme k-mean avec k=3

Résultats pour k = 3

19:42:31 - SimpleKMeans

```
--- Run information ---

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     cereals-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    77
Attributes:   16
              calories
              protein
              fat
              sodium
              fiber
              carbo
              sugars
              potass
              vitamins
              shelf
              weight
              cups
              rating
Ignored:
              name
              mfr
              type
Test mode:    evaluate on training data


--- Clustering model (full training set) ---


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 34.780935251125906

Initial starting points (random):

Cluster 0: 0.454545,0.2,0,0.90625,0.071429,0.916667,0.1875,0.108761,0.25,0,0.5,0.6,45.863324
Cluster 1: 0.181818,0.6,0.2,0.40625,0.714286,0.25,0.4375,0.848943,0.25,1,0.5,0.064,68.402973
Cluster 2: 0.545455,0,0.2,0.5625,0,0.541667,0.875,0.199396,0.25,0.5,0.5,0.6,22.396513
```

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#		
Attribute	Full Data	0	1	2
	(77.0)	(24.0)	(16.0)	(37.0)
calories	0.5171	0.4735	0.3466	0.6192
protein	0.3091	0.3333	0.4	0.2541
fat	0.2026	0.1167	0.175	0.2703
sodium	0.499	0.5052	0.3613	0.5545
fiber	0.1537	0.1122	0.2969	0.1187
carbo	0.6499	0.717	0.5508	0.6492
sugars	0.4951	0.3411	0.3398	0.6622
potass	0.2933	0.2213	0.4326	0.2797
vitamins	0.2825	0.1979	0.2188	0.3649
shelf	0.6039	0.0833	0.9688	0.7838
weight	0.5296	0.4929	0.4375	0.5932
cups	0.4568	0.548	0.2795	0.4744
rating	42.6657	47.5374	55.8284	33.8137

Time taken to build model (full training data) : 0.01 seconds

--- Model and evaluation on training set ---

Clustered Instances

0 24 (31%)
1 16 (21%)
2 37 (48%)

Répartition :

- Cluster 0 : 24 instances (31%)
- Cluster 1 : 16 instances (21%)
- Cluster 2 : 37 instances (48%)

Analyse des clusters (k = 3)

Cluster 0 (31%)

- calories = 0.4735
- protein = 0.3333
- fat = 0.1167
- sodium = 0.5052
- fiber = 0.1122
- carbo = 0.717
- sugars = 0.3411
- potass = 0.2213
- vitamins = 0.1979
- shelf = 0.0833
- weight = 0.4929
- cups = 0.548
- rating = 47.5374

Interprétation:

Ce groupe représente le cœur du marché. Ce sont des céréales équilibrées, principalement axées sur les glucides.

- Analyse des nutriments : Avec le taux de carbo le plus élevé (0.717), ce cluster privilégie l'apport énergétique. Le taux de sucre est modéré (0.34) et les graisses sont faibles (0.11).
- Positionnement : Le score shelf (0.08) est très bas, indiquant que ces produits sont systématiquement placés sur l'étagère du bas. Ce sont les produits familiaux de base.
- Verdict (Rating 47.54) : Un score très honorable qui montre que la simplicité (peu de gras, sucre contrôlé) paie au niveau nutritionnel.

Conclusion :

Le Cluster 0 regroupe les céréales dites 'énergétiques’. Elles offrent un apport maximal en glucides complexes avec très peu de matières grasses. Leur positionnement en rayon bas et leur score de qualité intermédiaire (Rating 47.54) en font le choix standard des familles cherchant un équilibre entre coût et énergie.

Cluster 1 (21%)

- calories = 0.3466
- protein = 0.4
- fat = 0.175
- sodium = 0.3613
- fiber = 0.2969
- carbo = 0.5508
- sugars = 0.3398
- potass = 0.4326
- vitamins = 0.2188
- shelf = 0.9688
- weight = 0.4375
- cups = 0.2795
- rating = 55.8284

Interprétation:

C'est le segment "haut de gamme" du point de vue de la santé. C'est ici que l'algorithme a isolé les meilleurs ingrédients.

- Analyse des nutriments : Ce cluster possède le moins de calories (0.346) et le plus de fibres (0.297). Il est également riche en protéines (0.4) et en potassium (0.43).
- Volume : Le score cups (0.27) est le plus bas, ce qui signifie que ce sont des céréales denses (une petite portion suffit).
- Verdict (Rating 55.83) : C'est le cluster champion en termes de qualité nutritionnelle.

Conclusion :

Le Cluster 1 représente le segment 'Bien-être et Santé'. Il se distingue par la densité nutritionnelle la plus élevée : riche en fibres et en protéines, tout en étant le plus pauvre en calories. Avec le meilleur Rating (55.83), il identifie les produits recommandés pour une alimentation diététique.

Cluster 2 (48%)

- calories = 0.6192
- protein = 0.2541
- fat = 0.2703
- sodium = 0.5545
- fiber = 0.1187

- carbo = 0.6492
- sugars = 0.6622
- potass = 0.2797
- vitamins = 0.3649
- shelf = 0.7838
- weight = 0.5932
- cups = 0.4744
- rating = 33.8137

Interprétation:

C'est le groupe le plus large (presque la moitié des céréales du dataset). Il regroupe tout ce qui est plus riche et plus transformé.

- Analyse des nutriments : Les calories (0.619), les sucres (0.662) et les graisses (0.27) y sont à leurs niveaux les plus hauts. L'apport en vitamines ajoutées est aussi le plus élevé (0.36), souvent pour compenser la perte de nutriments naturels.
- Positionnement : Une étagère élevée (0.78), typique des produits très marketés (marques célèbres, céréales pour enfants).
- Verdict (Rating 33.81) : C'est la note la plus basse, pénalisée par l'excès de sucre et de calories.

Conclusion:

Le Cluster 2 regroupe les céréales dites “Moins saines”. Occupant près de la moitié du marché, ce segment privilégie le goût avec des taux élevés de sucre et de lipides. En conséquence, il affiche la qualité nutritionnelle la plus faible (Rating 33.81), marquant une séparation nette avec les produits de santé."

V-Comparaison finale et choix de la solution la plus satisfaisante

Comparaison quantitative

WCSS: Within cluster sum of squared errors (somme des carrés des erreurs au sein du cluster)

- k = 5
    - WCSS = 25.71312405440395
    - Clusters plus compacts
    - Meilleure homogénéité intra-cluster
- k = 3
    - WCSS = 34.788935251125906
    - Clusters plus dispersés
    - Perte d'information nutritionnelle

Comparaison qualitative

k = 5

- Séparation plus fine
- Groupes nutritionnellement cohérents
- Interprétation détaillée par profil
- Segmentation claire :
    - céréales très sucrées
    - céréales santé riches en fibres
    - céréales enrichies en vitamines
    - céréales équilibrées
    - céréales très caloriques

k = 3

- Groupes trop larges
- Mélange de profils nutritionnels
- Analyse moins précise
- Moins adapté à une segmentation nutritionnelle

Solution la plus satisfaisante:

Après étude des résultats fournis par l'algorithme K-Means sur le dataset des céréales, nous préconisons la solution à k = 5 comme étant la plus pertinente pour les raisons suivantes :

1. Une meilleure finesse de segmentation (Précision)

Alors que la solution à k = 3 se contente de séparer le marché en trois blocs massifs (Santé, Moyen, Moins sains), la solution à k = 5 permet d'isoler des comportements nutritionnels plus subtils :

- Elle distingue les céréales très sucrées (Cluster 2) des céréales denses et grasses (Cluster 4). Dans la solution à k=3, ces deux catégories sont fusionnées dans le même groupe, ce qui masque leurs différences structurelles.
- Elle isole un "pôle d'excellence" (Cluster 1, 9% du dataset) qui présente des taux de fibres et de potassium exceptionnels, lesquels sont dilués dans la solution à k=3.

2. Une corrélation plus précise avec la variable Rating

L'objectif de l'analyse était de comprendre ce qui détermine la qualité nutritionnelle.

- Avec k = 5, nous observons une hiérarchie très claire des Ratings (64.8, 53.0, 44.1, 36.3, 28.3).
- Cette décomposition montre que le score de qualité ne chute pas seulement à cause du sucre, mais aussi à cause de la densité calorique et du sodium, des nuances que seule la solution à 5 clusters met en évidence.

3. Cohérente avec la réalité du marché

La solution à k = 5 correspond plus fidèlement à la segmentation réelle en grande distribution :

1. Diététique (Cluster 1)
2. Énergétique et Économique (Cluster 0)
3. Moyen (Cluster 3)
4. Énergétique (Cluster 4)
5. Moins sains (Cluster 2)

Conclusion

Bien que la solution à k = 3 soit plus robuste statistiquement par la taille de ses segments, elle reste trop généraliste. La solution à k = 5 apporte une valeur explicative supérieure, transformant de simples données numériques en une véritable connaissance métier exploitable.

Bibliographie

Téléchargement du dataset (cereals.csv) https://www.kaggle.com/crawford/datasets?query=cereal

Conclusion

Au terme de cette étude réalisée sous WEKA, l'application de l'algorithme K-Means a permis de transformer des données brutes en une cartographie lisible du marché des céréales. La comparaison entre les deux modèles révèle des enseignements distincts :

- La solution à k=3 offre une vision simplifiée, séparant efficacement les produits « santé » des produits « moins sains ». Elle est idéale pour une aide à la décision rapide.
- La solution à k=5 s'est avérée être la plus satisfaisante. En effet, elle affine l'analyse en isolant des segments spécifiques, tels que les céréales énergétiques denses (riches en gras et calories) par rapport aux céréales purement sucrées. Cette précision permet de mieux comprendre la variabilité du Rating (score de qualité), qui ne dépend pas d'un seul facteur mais d'une combinaison complexe de nutriments.

En conclusion, ce clustering démontre que l'intelligence artificielle peut reproduire, et même affiner, les segmentations marketing et diététiques réelles. Ce travail pose les bases d'un système expert capable de classer automatiquement tout nouveau produit céréalier et d'en prédire la valeur nutritionnelle relative, ouvrant ainsi la voie à des outils d'assistance à la consommation plus intelligents.