

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



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



Filière : INFORMATIQUE DE
GESTION (Analyse Informatique
et Programmation)

Discipline:
INTRODUCTION A
L'INTELLIGENCE
ARTIFICIELLE

TP1:
SEGMENTATION ~ MACHINE LEARNING

Membres du Groupe 1

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

Professeur

Dr DAGBA Théophile Komlan

ANNÉE ACADEMIQUE

2025 – 2026

Sommaire

Table des matières

Page de garde	1
Sommaire	2
Table des matières.....	2
Introduction générale.....	3
I-Compréhension des données.....	3
I-1-Attributs utilisés dans l'analyse	3
I-2-Attributs ignorés.....	4
II-Prétraitement des données.....	4
III-1 Gestion des valeurs manquantes : <i>ReplaceMissingValues</i>	4
III-2 Normalisation des données : <i>Normalize</i>	4
IV-Implémentation dans WEKA	5
IV-1-Algorithme k-mean avec k=5	5
Résultats pour k = 5	5
IV-2-Algorithme k-mean avec k=3	9
Résultats pour k = 3	9
V-Comparaison finale et choix de la solution la plus satisfaisante	11
Solution la plus satisfaisante:	12
VI-Utilisation avec Python(K-means)	12
VI-1. Rappel du contexte et de la méthode.....	12
VI-2. Analyse des résultats pour k = 5	13
VI-3. Analyse des résultats pour k = 3	14
VI-4. Comparaison finale k = 5 vs k = 3	16
VI-5. Conclusion	16
Conclusion générale	17

Introduction générale

Avec l'augmentation des volumes de données disponibles, les techniques de fouille de données (data mining) jouent un rôle essentiel dans l'extraction d'informations pertinentes à partir de bases de données complexes. Parmi ces techniques, le clustering permet de regrouper automatiquement des données similaires sans connaissance préalable des classes.

Dans ce travail, nous nous intéressons au dataset « Cereals », disponible sur le site *Machine Learning Repository*, qui contient des informations nutritionnelles sur différentes céréales pour petit-déjeuner. Chaque céréale est décrite par plusieurs attributs quantitatifs tels que : calories, protéines, lipides, sodium, fibres, glucides, sucres, potassium, vitamines, durée de conservation (shelf), poids, contenance (cups) et note (rating).

L'objectif principal de ce TP est d'appliquer l'algorithme de clustering k-means afin d'identifier des groupes homogènes de céréales ayant des caractéristiques nutritionnelles proches. Pour cela, deux valeurs du nombre de clusters sont étudiées : $k = 5$ et $k = 3$.

L'analyse est réalisée dans un premier temps à l'aide du logiciel WEKA, puis reproduite en utilisant le langage de machine learning Python, afin de comparer les résultats obtenus et d'évaluer la solution la plus satisfaisante.

I-Compréhension des données

Le jeu de données utilisé contient 77 instances, 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

Résultats pour k = 5

```
18:16:40 - FM
--- Input information ---
Scheme: weka.clusterers.EuclideanDistanceClusterer -I 100 -N 5 -X 10 -max -1 -lrcv 1.0E-6 -litter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation: cereals-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances: 77
Attributes: 16
calories
protein
fat
sodium
fiber
carbs
sugars
potass
vitamins
shelff
weight
cups
rating
Ignored:
name
mfr
type
Test mode: evaluate on training data

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

FM
==

Number of clusters: 5
Number of iterations performed: 2

      Cluster
Attribute   0     1     2     3     4
          (0.12) (0.13) (0.27) (0.26) (0.22)

calories
mean       0.5735  0.2504  0.5584  0.4544  0.6741
std. dev.  0.1393  0.1915  0.0424  0.0761  0.1597

18:16:40 - FM
protein
mean       0.292  0.3909  0.1948  0.3997  0.4175
std. dev.  0.0997  0.1937  0.1174  0.260  0.0897

fat
mean       0.1356  0.063  0.2095  0.1196  0.4177
std. dev.  0.0936  0.0929  0.1591  0.1466  0.2221

sodium
mean       0.6582  0.3919  0.5357  0.4649  0.4779
std. dev.  0.1423  0.2709  0.1783  0.3672  0.1725

fiber
mean       0.106  0.3494  0.0374  0.1349  0.2295
std. dev.  0.0905  0.3063  0.0465  0.0975  0.0933

carbs
mean       0.0205  0.3463  0.1655  0.746  0.62
std. dev.  0.0938  0.172  0.069  0.2121  0.1329

sugars
mean       0.4023  0.260  0.3738  0.2343  0.6484
std. dev.  0.2225  0.1525  0.1272  0.1422  0.1636

potass
mean       0.2534  0.4736  0.1208  0.2461  0.4697
std. dev.  0.1731  0.3276  0.0586  0.1179  0.1405

vitamins
mean       0.7405  0.2311  0.25  0.1876  0.2349
std. dev.  0.3568  0.0992  0.2234  0.1082  0.0596

shelff
mean       1     1     0.3809  0.1525  0.9689
std. dev.  0.4163  0.0904  0.2629  0.234  0.1207

weight
mean       0.5973  0.4922  0.5  0.4915  0.66
std. dev.  0.1717  0.1903  0  0.037  0.173

cups
mean       0.5332  0.3934  0.5002  0.5482  0.3437
std. dev.  0.0931  0.2261  0.1271  0.173  0.1510
```

```

rating
mean      40.1462 61.4907 28.576 52.636 30.2369
std. dev.  5.7277 11.7729 5.2499 10.4765 4.4706

Time taken to build model (full training data) : 0.09 seconds
--- Model and evaluation on training set ---
Clustered Instances
0      8 ( 10%)
1     12 ( 16%)
2     21 ( 27%)
3     20 ( 26%)
4     16 ( 21%)
log likelihood: 6.17144

```

Nombre total d'instances : **77 céréales**

Répartition :

- Cluster 0 : **8 (10 %)**
- Cluster 1 : **12 (16 %)**
- Cluster 2 : **21 (27 %)**
- Cluster 3 : **20 (26 %)**
- Cluster 4 : **16 (21 %)**

Cluster 0 (10 %)

- calories : **0.5735**
- protéines : **0.292**
- lipides : **0.1356**
- sodium : **0.6582**
- fibres : **0.106**
- glucides : **0.8205**
- sucres : **0.4023**
- potassium : **0.2534**
- vitamines : **0.7405**
- durée de conservation : **1**
- poids : **0.5873**
- contenance : **0.5332**
- note : **40.1462**

Conclusion: Céréales très riches en glucides et vitamines, avec un sodium élevé et une note moyenne. Profil nutritionnel dense mais pas optimal en termes d'évaluation globale.

Cluster 1 (16 %)

- calories : **0.2504**
- protéines : **0.3909**
- lipides : **0.063**
- sodium : **0.3819**
- fibres : **0.3494**
- glucides : **0.5463**
- sucres : **0.269**
- potassium : **0.4736**
- vitamines : **0.2011**
- durée de conservation : **1**
- poids : **0.4022**
- contenance : **0.3034**
- note : **61.4907**

Conclusion: Céréales faibles en calories et lipides, riches en fibres, protéines et potassium, avec la meilleure note globale. Profil nutritionnel le plus équilibré.

Cluster 2 (27 %)

- calories : **0.5584**
- protéines : **0.1048**
- lipides : **0.2095**
- sodium : **0.5357**
- fibres : **0.0374**
- glucides : **0.5655**
- sucres : **0.7738**
- potassium : **0.1288**
- vitamines : **0.25**
- durée de conservation : **0.3809**
- poids : **0.5**
- contenance : **0.5002**
- note : **28.576**

Conclusion: Céréales très riches en sucres, pauvres en fibres, protéines et potassium, avec la plus faible note. Céréales à faible qualité nutritionnelle.

Cluster 3 (26 %)

- calories : **0.4544**
- protéines : **0.3997**
- lipides : **0.1196**
- sodium : **0.4649**
- fibres : **0.1349**
- glucides : **0.746**
- sucres : **0.2343**
- potassium : **0.2461**
- vitamines : **0.1876**
- durée de conservation : **0.1525**
- poids : **0.4915**
- contenance : **0.5482**
- note : **52.636**

Conclusion: Céréales riches en protéines et glucides, faibles en sucres, avec une **bonne note**. Bon compromis nutritionnel.

Cluster 4 (21 %)

- calories : **0.6741**
- protéines : **0.4175**
- lipides : **0.4177**
- sodium : **0.4778**
- fibres : **0.2295**
- glucides : **0.61**
- sucres : **0.6484**
- potassium : **0.4697**
- vitamines : **0.2349**
- durée de conservation : **0.9689**
- poids : **0.66**
- contenance : **0.3437**
- note : **38.2369**

Conclusion: Céréales énergétiques, riches en **protéines, lipides et sucres**, avec une **note modérée**.
 Produits plus lourds et plus riches, mais moins bien évalués.

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

Résultats pour k = 3

```

18:18:00 -FM
--- Run information ---
Scheme: weka.clustering_EM -I 100 -N 3 -X 10 -max -1 -l -cv 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation: cereals-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Normalize-S1,0-90.0-weka.filters.unsupervised.attribute.Normalize-S1,0-90.0
Instances: 77
Attributes: 16
    calories
    protein
    fat
    sodium
    fiber
    carbs
    sugars
    potass
    vitamins
    shelf
    weight
    cups
    rating
Ignored:
    name
    air
    type
Test mode: evaluate on training data

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

18:18:00 -FM
Number of clusters: 3
Number of iterations performed: 1

      Cluster
Attribute   0     1     2
             (0.26) (0.27) (0.47)
calories
mean       0.5589  0.4554  0.3207
std. dev.  0.0432  0.0763  0.2411

18:18:00 -FM
protein
mean       0.1006  0.3969  0.3756
std. dev.  0.1174  0.2662  0.1404

fat
mean       0.203  0.1185  0.2498
std. dev.  0.156  0.1462  0.2302

sodium
mean       0.5319  0.4600  0.4997
std. dev.  0.1795  0.3666  0.2232

fiber
mean       0.036  0.134  0.2304
std. dev.  0.0006  0.0969  0.2016

carbs
mean       0.5631  0.7474  0.6482
std. dev.  0.069  0.2109  0.1708

sugars
mean       0.7818  0.2365  0.4015
std. dev.  0.1228  0.1435  0.2406

potass
mean       0.1327  0.2447  0.1102
std. dev.  0.0541  0.1176  0.2415

vitamins
mean       0.29  0.109  0.3533
std. dev.  0.0053  0.1074  0.2957

shelf
mean       0.3616  0.1657  0.0862
std. dev.  0.2402  0.3551  0.0819

weight
mean       0.5  0.4957  0.5675
std. dev.  0  0.0366  0.2096

cups
mean       0.5041  0.5473  0.3794
std. dev.  0.128  0.1721  0.1867

rating
mean       28.3506 52.4626 45.0629
std. dev.  5.2206 10.4081 12.7670

Time taken to build model (full training data) : 0.04 seconds
--- Model and evaluation on training set ---
Clustered instances
0     23 ( 30%)
1     20 ( 26%)
2     34 ( 44%)
Log likelihood: 5.0164

```

Répartition :

- Cluster 0 : **23 (30 %)**
- Cluster 1 : **20 (26 %)**
- Cluster 2 : **34 (44 %)**

Cluster 0 (30 %)

- calories : **0.5589**
- protéines : **0.1006**
- lipides : **0.203**
- sodium : **0.5319**
- fibres : **0.036**
- glucides : **0.5631**
- sucres : **0.7818**
- potassium : **0.1327**
- vitamines : **0.25**
- durée de conservation : **0.3616**
- poids : **0.5**
- contenance : **0.5041**
- note : **28.3508**

Conclusion: Groupe des céréales très sucrées, pauvres en fibres et protéines, avec la plus faible note.

Cluster 1 (26 %)

- calories : **0.4554**
- protéines : **0.3969**
- lipides : **0.1185**
- sodium : **0.4688**
- fibres : **0.134**
- glucides : **0.7474**
- sucres : **0.2365**
- potassium : **0.2447**

- vitamines : **0.189**
- durée de conservation : **0.1657**
- poids : **0.4917**
- contenance : **0.5473**
- note : **52.4828**

Conclusion: Céréales bien notées, riches en protéines et glucides, faibles en sucres.

Cluster 2 (44 %)

- calories : **0.5287**
- protéines : **0.3756**
- lipides : **0.2498**
- sodium : **0.4977**
- fibres : **0.2304**
- glucides : **0.6432**
- sucres : **0.4815**
- potassium : **0.4102**
- vitamines : **0.3533**
- durée de conservation : **0.9862**
- poids : **0.5675**
- contenance : **0.3794**
- note : **45.0929**

Conclusion: Groupe hétérogène, valeurs moyennes sur presque tous les attributs.

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

- **k = 5 :**
 - Log-likelihood k=5 = 6.17
 - Séparation **plus fine**
 - Groupes **nutritionnellement cohérents**

- Interprétation nutritionnelle plus fine et détaillée par profil permettant une vraie segmentation marketing et nutritionnelle
- Groupes plus homogènes (Chaque cluster correspond à un profil nutritionnel précis; très sucré, santé, enrichi, équilibré, très calorique)
- **k = 3 :**
 - Log-likelihood k=3 = 5.01
 - Groupes **trop larges**
 - Mélange de profils nutritionnels différents
 - Moins précis pour l'analyse
 - Avec seulement 3 clusters, plusieurs profils nutritionnels différents sont regroupés ensemble, ce qui réduit la précision de l'analyse.

Solution la plus satisfaisante:

La segmentation avec k=5 est plus satisfaisante car elle permet d'identifier des profils nutritionnels distincts et cohérents : céréales très sucrées, céréales santé riches en fibres, céréales enrichies en vitamines, céréales équilibrées et céréales très caloriques. Le modèle k=5 présente un meilleur log-likelihood et offre une interprétation plus précise que k=3, qui regroupe des profils hétérogènes.

VI-Utilisation avec Python(K-means)

VI-1. Rappel du contexte et de la méthode

Après avoir appliqué l'algorithme **K-means** sous WEKA, nous avons reproduit la même démarche avec **Python**, en utilisant le dataset **Cereals**.

Les étapes suivies sont identiques à WEKA :

- Suppression des attributs non numériques (*name, mfr, type*)
- Remplacement des valeurs manquantes par la **moyenne**
- Normalisation des données entre **0 et 1**
- Application de **K-means** pour :
 - **k = 5**
 - **k = 3**

Les attributs étudiés sont :

calories, protéines, lipides, sodium, fibres, glucides, sucres, potassium, vitamines, durée de conservation (shelf), poids, contenance (cups), note (rating).

VI-2. Analyse des résultats pour k = 5

Répartition des clusters

- Cluster 0 : 26 céréales (33,8 %)
- Cluster 1 : 13 céréales (16,9 %)
- Cluster 2 : 18 céréales (23,4 %)
- Cluster 3 : 11 céréales (14,3 %)
- Cluster 4 : 9 céréales (11,7 %)

La répartition est relativement équilibrée, ce qui indique une **bonne segmentation** du dataset.

Interprétation des clusters (k = 5)

Cluster 0 – Céréales riches et relativement bien notées

- Calories élevées (0.587)
- Protéines élevées (0.423)
- Fibres modérées (0.255)
- Sucres assez élevés (0.582)
- Potassium élevé (0.469)
- Bonne note moyenne (0.323)

Ce groupe correspond à des céréales **énergétiques**, assez complètes sur le plan nutritionnel.

Cluster 1 – Céréales très riches en glucides et sodium

- Sodium très élevé (0.764)
- Glucides élevés (0.790)
- Fibres faibles (0.102)
- Sucres modérés (0.317)
- Note moyenne faible (0.342)

Ce sont des céréales **peu équilibrées**, riches en sel et glucides, moins bien notées.

Cluster 2 – Céréales très sucrées et peu protéinées

- Sucres très élevés (0.802)
- Protéines très faibles (0.078)
- Fibres très faibles (0.030)

- Note la plus faible (0.128)

Ce groupe représente des céréales **très sucrées**, peu nutritives, souvent destinées aux enfants.

Cluster 3 – Céréales légères et bien notées

- Calories faibles (0.289)
- Lipides très faibles (0.055)
- Sodium très faible (0.067)
- Sucres faibles (0.142)
- Note élevée (0.628)

Ce sont des céréales **diététiques**, légères et bien appréciées.

Cluster 4 – Céréales enrichies en vitamines

- Vitamines très élevées (0.667)
- Glucides élevés (0.838)
- Sodium élevé (0.656)
- Note moyenne (0.334)

Ce cluster correspond à des céréales **fortement enrichies**, mais pas forcément les mieux notées.

Qualité du modèle (k = 5)

- **Silhouette score = 0.2578**

Ce score indique une **séparation correcte** entre les groupes.

VI-3. Analyse des résultats pour k = 3

Répartition des clusters

- Cluster 0 : 36 céréales (46,8 %)
- Cluster 1 : 11 céréales (14,3 %)
- Cluster 2 : 30 céréales (39,0 %)

La répartition est plus grossière, avec des groupes plus larges.

Interprétation des clusters (k = 3)

Cluster 0 – Céréales globalement équilibrées

- Calories élevées (0.556)
- Protéines élevées (0.394)
- Fibres élevées (0.236)
- Sucres modérés (0.502)
- Note moyenne (0.345)

Groupe général de céréales **assez équilibrées**.

Cluster 1 – Céréales légères et bien notées

- Calories faibles (0.339)
- Sodium très faible (0.092)
- Sucres faibles (0.159)
- Note la plus élevée (0.579)

Céréales **diététiques**, proches du cluster 3 de k=5.

Cluster 2 – Céréales riches et sucrées

- Sodium élevé (0.615)
- Sucres élevés (0.610)
- Fibres faibles (0.060)
- Note faible (0.209)

Céréales **moins saines**, riches en sucre et sel.

Qualité du modèle (k = 3)

- **Silhouette score = 0.2556**

Ce score est légèrement inférieur à celui obtenu pour k=5.

VI-4. Comparaison finale k = 5 vs k = 3

Critère	k = 5	k = 3
Nombre de groupes	Plus détaillé	Plus général
Interprétation nutritionnelle	Très fine	Plus grossière
Silhouette score	0.2578	0.2556
Pertinence globale	Meilleure	Bonne

VI-5. Conclusion

La solution **k = 5** est la plus satisfaisante car :

- Elle permet une **meilleure distinction nutritionnelle** entre les céréales
- Les groupes sont plus **homogènes**
- Le **silhouette score est légèrement supérieur**
- L'interprétation est plus riche et plus proche de celle obtenue avec **WEKA**

Nous retenons donc k = 5 comme la meilleure solution, aussi bien sous WEKA que sous Python.

Conclusion générale

Dans ce travail, nous avons appliqué l'algorithme de **k-means** sur le dataset *Cereals* après une phase essentielle de **prétraitement des données**, comprenant :

- le remplacement des valeurs manquantes par la moyenne (*ReplaceMissingValues*),
- la normalisation des données entre 0 et 1 (*Normalize*), afin de garantir une contribution équitable de tous les attributs au processus de clustering.

Les expérimentations ont été menées pour **k = 5** et **k = 3**, aussi bien avec **WEKA** qu'avec **Python**, en utilisant les mêmes attributs nutritionnels.

Les résultats montrent que :

- la solution **k = 5** permet une **segmentation plus fine** des céréales, en distinguant clairement plusieurs profils nutritionnels (céréales riches en sucres, céréales plus équilibrées, céréales riches en fibres ou en vitamines, etc.) ;
- la solution **k = 3**, bien que plus simple, regroupe des profils parfois trop hétérogènes et perd en précision descriptive.

De plus, l'évaluation quantitative à l'aide du **silhouette score** en Python montre que **k = 5 obtient un score légèrement supérieur à k = 3**, indiquant une meilleure séparation des clusters. Ces observations sont cohérentes avec les résultats obtenus sous WEKA, ce qui valide la pertinence de l'implémentation Python.

Pour conclure, choisir **5 groupes (k=5)** est la meilleure solution. Cela permet de créer des catégories de céréales précises et faciles à comprendre selon leur nutrition (les plus sucrées, les plus équilibrées, etc.).

Ce travail montre que le regroupement automatique (**clustering**) est très utile pour analyser des données. Enfin, nous avons vu que **WEKA** et **Python** travaillent bien ensemble : WEKA est pratique pour tester rapidement, et Python est puissant pour personnaliser l'analyse.