Projet SY09 2025

Bengriche Tidiane, Tareq Derdaki, Ruoyang Wang

12/06/2025

Introduction

Le jeu de données retenu contient des **informations nutritionnelles** détaillées pour différents aliments/plats. Il comporte **3454 lignes**, chacune correspondant à un aliment/plat, et **11 variables** représentant la teneur en 11 nutriments. Le jeu de données est accessible ici.

1 ACP

Après avoir normalisé les données et supprimé les valeurs aberrantes, nous avons appliqué une analyse en composantes principales (PCA) afin de réduire la dimensionnalité et de faciliter l'analyse du jeu de données. Après avoir calculé les composantes principales, on se rend compte que l'inertie est expliquée globalement de manière équitable (sauf pour le premier axe) entre tous les axes (Figure 11 en Annexes). La première composante représente globalement un axe de densité énergétique (valeur calorique, protéines, lipides, glucides), tandis que la deuxième met en évidence un contraste entre aliments riches en glucides et sucres versus aliments riches en protéines et lipides. En ne gardant que ces **2 premières composantes**, on obtient le graphe suivant en effectuant un regroupement hiérarchique :

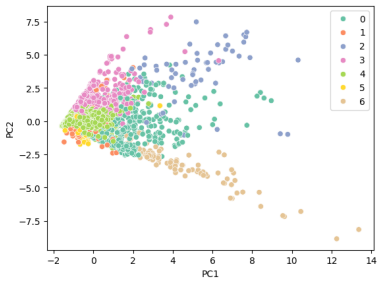

Figure 1 – Résultat de l'ACP

ici, nous avons essayé avec 7 classes. Malheureusement, après examen des classes proposées par l'algorithme, nous n'arrivions pas à en dégager des classes qui correspondent à la réalité.

2 AFTD

L'ACP n'ayant pas permis de dégager une structure claire parmi nos variables, nous nous sommes tournés vers l'AFTD qui offre une bien meilleure représentation des distances entre les individus. Dans l'optique de réaliser un clustering un peu plus tard, l'AFTD répond parfaitement à nos besoins.

Nous pouvons observer la projection de nos données dans un espace à deux dimensions obtenue par l'AFTD. En théorie, cette représentation devrait nous permettre d'observer les distances entre des couples ou groupes d'aliments. Cependant, le nombre d'individus étant très élevé, il n'est pas possible de distinguer un aliment particulier des autres grâce à des étiquettes ou d'autres méthodes.

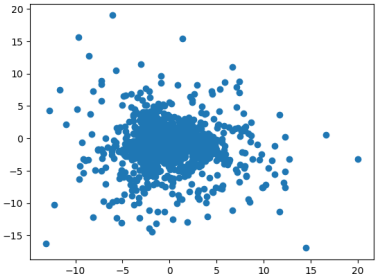

Figure 2 – AFTD

Pour le moment, il est même assez complexe de distinguer des groupes car on semble uniquement disposer d'un gros tas d'individus rassemblés au même endroit. Toutefois, cette observation préliminaire confirme la pertinence d'un approfondissement en termes de clustering, que nous développerons dans la section suivante.

3 Clustering et Classification Hiérarchique

Suite à notre analyse par AFTD, nous avons procédé à une classification hiérarchique sur cette même AFTD

afin d'identifier des groupes d'aliments présentant des similarités au niveau de leurs profils nutritionnels.

3.1 Première approche par dendrogramme

Dans un premier temps, nous avons réalisé une classification hiérarchique ascendante avec lien minimal, que nous avons représentée grâce à un dendrogramme. Malheureusement, la structure arborescente obtenue ne nous a pas permis de distinguer clairement des groupes cohérents d'aliments (Figure 12 en Annexes). Face à cette limitation, nous avons décidé d'adopter une approche différente en combinant le clustering avec la représentation bidimensionnelle fournie par l'AFTD précédemment réalisée.

3.2 Tentative de clustering en 6 groupes

Pour commencer, nous avons tenté de segmenter notre jeu de données en 6 clusters, en nous inspirant de la classification nutritionnelle proposée par l'AMELI qui distingue : **Fruits et légumes** ; **Produits laitiers** ; **Viandes, poissons et fruits de mer** ; **Féculents et légumes secs** ; **Matières grasses** ; **Produits sucrés**.

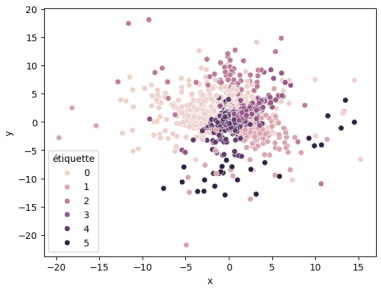

FIGURE 3 – Clustering en 6 groupes

Cette première tentative s'est avérée peu concluante, avec des frontières très floues entre les différents groupes.

3.3 Optimisation avec 4 clusters et Interprétation des résultats

Face à ces résultats, nous avons opté pour une réduction du nombre de clusters à **4**, cherchant ainsi à obtenir des groupes plus cohérents et mieux délimités. Évidemment, ici, nous ne nous attendons pas à retrouver les mêmes catégories que celles proposées par l'AMELI, mais plutôt à obtenir une catégorisation plus large se basant sur d'autres critères comme, par exemple, des individus à tendance grasse ou sucrée.

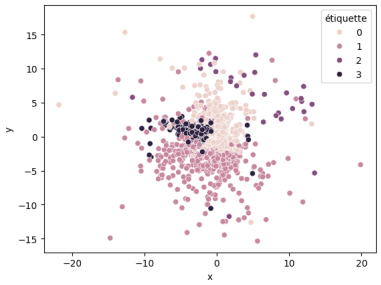

FIGURE 4 – Clustering en 4 groupes

Pour interpréter ces résultats, nous avons associé à chaque aliment son cluster d'appartenance et exporté ces données au format CSV afin d'effectuer une analyse plus simple sous Excel.

L'analyse des caractéristiques nutritionnelles dominantes au sein de chaque cluster nous a permis d'établir la typologie suivante :

- **Groupe 0** : Aliments relativement gras sans autre caractéristique nutritionnelle marquante.
- **Groupe 1** : Aliments à fort apport calorique, généralement riches en matières grasses mais également en protéines et en calcium.
- **Groupe 2** : Aliments caractérisés par une très forte teneur en sodium.
- **Groupe 3** : Aliments riches en glucides complexes et simples, avec une proportion importante de sucres.

Cette catégorisation en quatre groupes offre une vision plus claire et plus exploitable de la structure de notre jeu de données que la tentative initiale avec six clusters, tout en conservant une cohérence nutritionnelle significative.

4 K-means

Sur la base du prétraitement effectué précédemment, nous avons poursuivi l'analyse en appliquant l'algorithme K-means à nos données. Étant donné la complexité du jeu de données complet et des premiers résultats peu concluants, nous avons adopté une approche exploratoire en constituant successivement plusieurs sous-ensembles. À chaque étape, la construction du sous-ensemble et le choix des catégories, **attribuées manuellement aux aliments**, ont été ajustés en fonction des limites identifiées lors de l'itération précédente. Trois sous-ensembles ont ainsi été constitués au cours de ce processus :

— Un **premier** ensemble comprenant 209 échantillons extraits de six grandes catégories du jeu de données : 1 - produits sucrés, 2 - matières grasses, 3 - produits laitiers, 4 - féculents, 5 - fruits et légumes, 6 - viandes et poissons.
— Un **second** ensemble centré uniquement sur les catégories : 1 - poissons, 2 - fruit, 3 - légumes, 4 - viandes, ainsi qu'un petit nombre d'aliments classés dans 0 - autres, soit 710 échantillons au total.
— Un **troisième** ensemble dérivé du deuxième, dans lequel les catégories ont été fusionnées en trois groupes : 0 - autres, 1 - poissons et viandes, 2 - fruits et légumes.

Pour chacun de ces ensembles, nous avons effectué une série de clustering **K-means**, en testant deux stratégies d'initialisation ("random" et "k-means++") et plusieurs graines aléatoires. Les résultats ont été visualisés en deux dimensions grâce à l'analyse en composantes principales (ACP), en les comparant visuellement aux catégories annotées. Parallèlement, nous avons analysé la qualité des clusters obtenus à travers deux indicateurs :l'inertie intra-cluster,et l'Adjusted Rand Index (ARI), qui mesure la cohérence entre les clusters et les catégories réelles. Nous avons identifié, pour chaque cas, la configuration ayant produit le meilleur ARI.

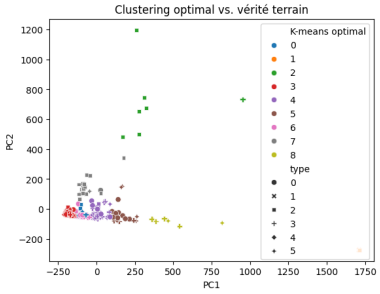

FIGURE 5 – Clustering optimal vs. vérité terrain

L'analyse a montré que, comme attendu, le deuxième jeu de données présente une inertie intra-cluster nettement plus faible, ce qui indique des groupes plus compacts. En revanche, son ARI est le plus bas des trois (au maximum 0,17), tandis que celui du premier ensemble atteint 0,25. Mais ce meilleur score est obtenu avec 9 clusters (Figure 5), alors que les catégories réelles sont au nombre de 6. Cela suggère que certaines catégories possèdent une structure interne complexe dans l'espace des variables, et que K-means a eu tendance à fragmenter ces catégories pour mieux minimiser l'inertie géométrique.

Le troisième ensemble a donné les meilleurs résultats globaux : la correspondance visuelle entre clusters et

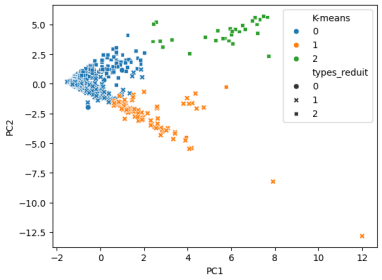

FIGURE 6 – Correspondance clusters/catégories

catégories est satisfaisante (Figure 6), et l'ARI maximal atteint 0,27, avec un nombre de clusters correspondant exactement au nombre de catégories (3). De manière générale, les ARI restent faibles dans les trois cas. Les graphiques "ARI en fonction de K et de la stratégie d'initialisation" (Figure 7) mettent en évidence une forte instabilité des résultats selon les graines aléatoires : pour une même configuration, les ARI varient fortement. Cela indique que l'algorithme K-means n'est pas très robuste dans ce contexte.

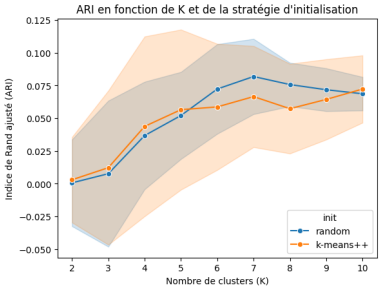

FIGURE 7 – ARI random/k-means++

Cette instabilité s'explique par la distribution des données : les variables présentent peu de corrélation et des répartitions très inégales. Dans le plan ACP, les points appartenant à une même catégorie sont alignés selon des directions allongées, et ces lignes issues de différentes catégories se rejoignent près de l'origine, formant une structure radiale. Cela va à l'encontre des hypothèses implicites de K-means (groupes sphériques, isotropes, de taille comparable).

5 Apprentissage supervisé

Pour poursuivre notre étude sur le jeu de données, il était logique de passer à de l'**apprentissage supervisé** sur ce dernier. L'objectif final de cette partie est d'utiliser des méthodes de machine learning (k-nn ou k-ppv, théorie Bayésienne de la décision, ADL, ADQ, etc.)

afin de pouvoir prédire, à partir des valeurs nutritionnelles d'un aliment, son appartenance à une catégorie d'aliments. La première méthode que nous utiliserons est évidemment celle qui est la plus connue, et celle qui fonctionne globalement le mieux : l'algorithme des k plus proches voisins (abrégé k-ppv ici).

5.1 Préparation du jeu d'apprentissage

La première étape pour pouvoir utiliser l'algorithme est de mettre en place un jeu de données dit d'apprentissage (ou d'entraînement). Pour ce faire, nous avons manuellement réalisé une sélection d'aliments selon des catégories bien distinctes (les effectifs dans chaque catégorie étant de taille similaire, entre 35 et 38 individus par catégorie). Le choix de prendre des catégories bien distinctes fait suite à notre précédente observation concluant que notre jeu de donnée était très regroupé et que de nombreux aliments ou plats (de catégories pourtant logiquement différentes) se confondaient. Les catégories d'aliments choisis ont été les suivantes (ces dernières étant toujours basés sur les principaux groupes établis par l'AMELI) :

— **Catégorie 1** : Produits sucrés
— **Catégorie 2** : Matières grasses
— **Catégorie 3** : Produits laitiers
— **Catégorie 4** : Féculents
— **Catégorie 5** : Fruits et légumes
— **Catégorie 6** : Viandes et poissons

Pour chaque individu, nous avons attribué sa catégorie avec le numéro correspondant dans une nouvelle colonne "type".

5.2 Analyse préliminaire et exclusion des matières grasses

Cependant, après avoir observé de manière simplifiée le jeu d'apprentissage grâce à une ACP, nous nous sommes rendus compte que les matières grasses étaient extrêmement isolées des autres données et que la variance intra-classe était très faible - les points de cette classe étaient très proches les uns des autres. Face à cette observation, nous avons pris la décision d'écarter cette catégorie pour réaliser notre apprentissage. Cette décision s'appuie sur deux arguments principaux :

— d'un côté, nous sommes quasi sûrs que toutes les matières grasses seraient correctement estimées par l'algorithme du fait de leur homogénéité nutritionnelle très différentiable des autres catégories.
— de l'autre côté, les écarter nous permet une meilleure observation et discrimination des autres catégories d'aliments, dont les profils nutritionnels présentent davantage de nuances et de recoupements.

5.3 Application de l'algorithme des k-ppv

5.3.1 Recherche du paramètre k optimal

Une fois les étapes préliminaires énoncées à l'instant accomplies, nous avons pu nous concentrer sur l'algorithme des k-ppv en lui-même. La première étape consistait à déterminer la valeur optimale du paramètre k, c'est-à-dire celle qui nous permettrait d'obtenir le meilleur taux de bonne classification des nouveaux individus. Pour cela, nous avons utilisé la méthode de validation croisée. Selon le nombre de plis choisi pour la validation croisée, nous obtenions des résultats légèrement différents : soit k=1, soit k=3 comme valeur optimale. Ci-dessus, nous pouvons observer le score obtenu pour 10 plis qui est maximisé en k=1 (pour cv=15, nous obtenons k=3) :

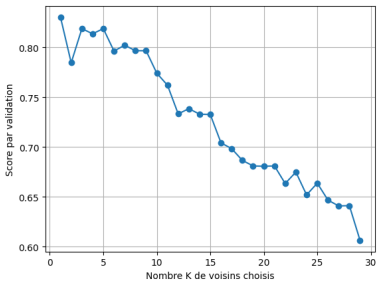

FIGURE 8 – Score par rapport aux nombre K de voisins

Afin d'obtenir les résultats les plus nuancés possibles, nous avons pris la décision d'observer, par la suite, les résultats pour les deux valeurs de k envisagées.

5.3.2 Exécution de l'algorithme

Une fois le paramètre k optimal déterminé, tous les éléments étaient en place pour lancer l'algorithme de classification sur notre jeu de données de test composé au total de 50 aliments et plats.

5.3.3 Analyse des résultats

L'exécution de l'algorithme aura permis d'observer des tendances très intéressantes, et pour la plupart, assez logiques. Tout d'abord, nous avons pu observer les taux de bonne classification suivants selon les valeurs de k choisies :

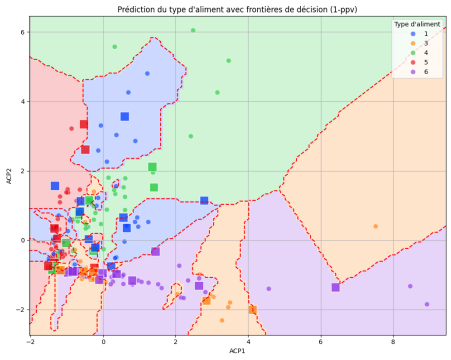

FIGURE 9 – Prédiction (1-ppv)

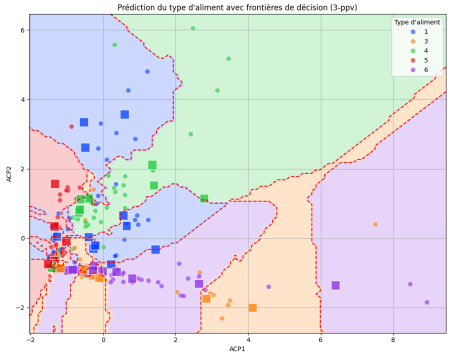

FIGURE 10 – Prédiction (3-ppv)

— k = 1 → 72%
— k = 3 → 56%

Pour la suite du rapport, nous nous concentrerons donc sur les observations que nous pouvons tirer du cas où **k=1**. Pourquoi k=1 ? On peut supposer que le jeu d'apprentissage est encore trop petit par rapport au nombre de colonnes et que l'algorithme tente de "compenser" ce manque de connaissances par une forme de surapprentissage et donc tend vers k=1 même si les frontières de décisions semblent proposer une certaine marge (dans la plupart des cas) autour des points. Dans le cas des classifications que nous considérons comme des erreurs claires, nous reconnaissons pour la très grande majorité de ces dernières, 2 schémas possibles plus ou moins compréhensibles :

— Des viandes qui sont catégorisées en produits laitiers → ces deux catégories sont nutritionnellement similaires car elles fournissent en grande quantité des protéines, du calcium (apport léger pour la viande) et des vitamines du groupe B qui sont deux colonnes importantes de notre jeu de données. On remarque d'ailleurs que ces catégories sont proches du point de vue de leur origine, puisqu'elles sont toutes deux de provenance animale.
— Des plats équilibrés qui sont catégorisés en produits sucrés → Les plats équilibrés ou du moins composés d'aliments "opposés" qui font que le plat est globalement "neutre" d'un point de vue nutritif terminent logiquement dans la catégorie qui est aussi la plus "neutre" (en effet, les produits sucrés n'ont quasiment aucun apport à part le sucre qu'ils contiennent).

Nous pouvons expliquer cela par le fait que le jeu d'apprentissage présente une taille insuffisante, ce qui peut conduire à des erreurs de classification pour les aliments situés en périphérie des catégories, car l'algorithme n'a pas été exposé à ces cas limites lors de l'entraînement.

En dehors de cela, on retrouve des aliments comme les yaourts au chocolat qui se retrouvent en produit sucré, ce qui reste compréhensible même si, à notre avis, l'humain aurait plutôt tendance à le catégoriser comme produit laitier. Nous considérerons ce genre de catégorisation comme correct étant donné qu'il peut représenter la meilleure option selon la perspective.

Nous pouvons donc en venir à la conclusion que le taux de bonne classification de l'algorithme pourrait être très satisfaisant pour des aliments qui appartiennent clairement aux catégories que nous avons mises en place, ce qui est finalement assez logique.

Conclusion

Les méthodes étudiées tout au long du semestre nous ont permis d'analyser notre jeu de données et d'en tirer des conclusions sur l'efficacité des algorithmes ainsi que sur le dataset lui-même :

— Premièrement, le choix du dataset n'a pas toujours permis d'identifier clairement des clusters dans certains cas. Il a souvent été nécessaire de réduire le nombre d'individus afin d'obtenir de meilleurs résultats. Cela s'explique par le grand nombre d'individus présents dans le dataset, mais aussi par la nature de ce dernier. En effet, bien que ce jeu de données contienne de nombreux aliments de base, peu ou pas transformés (pomme, chocolat, etc.), il intègre également de nombreux plats combinant plusieurs types d'aliments. Cette complexité rendait difficile l'identification de groupes cohérents.
— Malgré cette difficulté, l'apprentissage supervisé a donné de très bons résultats, notamment parce que l'affectation aux classes était réalisée manuellement. Par ailleurs, cette approche nous a permis de mettre en évidence les liens nutritifs entre cer-

tains types de nourriture.

D'un point de vue pratique, on pourrait imaginer que l'utilisation de ces algorithmes pourrait permettre de situer un plat par rapport aux autres, d'identifier s'il est trop salé ou sucré, et de vérifier l'équilibre entre les différents apports nutritifs, simplement en le comparant aux autres plats.

Annexes

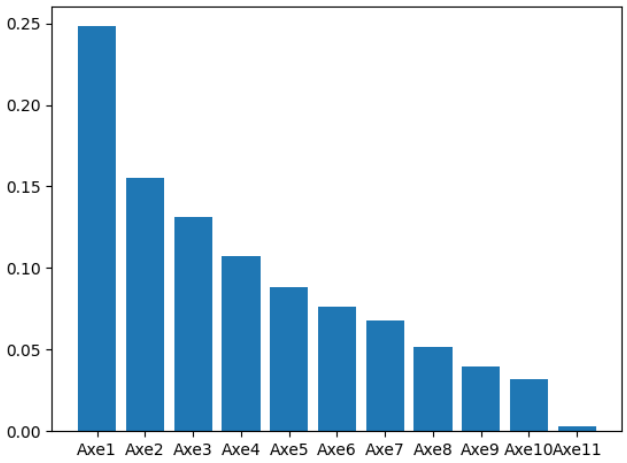

FIGURE 11 – Inertie expliquée

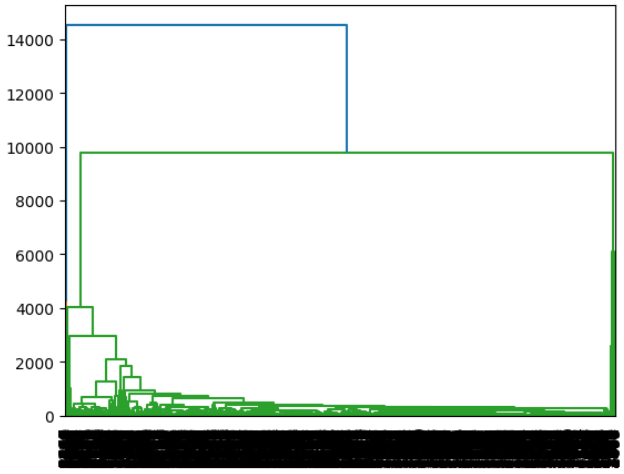

FIGURE 12 – Dendogramme