

Faculté des sciences

Identifying heterogeneity among cannabis users : a segmentation approach

Auteure : Keline Laure Meniodem Deliota

Promotrices : Pr. Catherine Legrand (UCLouvain / LSBA)

Dr. Elena Damian (Sciensano)

Lecteur : Pr. Christian Hafner

Année académique 2022-2023

Master[120] en science des données, orientation statistique, à finalité spécialisée

Table des matières

Remerciements	iv
Résumé	v
Abstract	vii
Liste des figures	ix
Liste des tableaux	xi
Liste des abréviations	xii
Introduction	1
1 Données et méthodologie utilisées dans cette étude	3
1.1 Données	3
1.1.1 Présentation des données	3
1.1.2 Variables utilisées dans cette étude	3
1.2 Méthodologie utilisée dans cette étude	6
1.2.1 Méthodes utilisées pour identifier les différents clusters	6
1.2.2 Méthodes utilisées pour caractériser les clusters identifiés	7
1.2.3 Méthodes utilisées pour la comparaison des profils observés sur la période 2015 à 2020 à ceux observés sur l'année 2021	7
2 Concepts théoriques et état de l'art	9
2.1 Analyse factorielle des données mixtes (AFDM)	10
2.1.1 Principe de l'AFDM	10
2.1.2 Intérêt de l'AFDM comme étape préalable à la classification	11
2.2 Quelques méthodes de clustering	12
2.2.1 Mesures de similarité	12
2.2.2 inertie intra-cluster et inter-cluster	15
2.2.3 Choix du nombre optimal de clusters	16
Technique du coude	17
Score de silhouette	17
2.2.4 K-means	19
Principe de base	20
Mesure de similarité	21
Choix du nombre de clusters	21

	Avantages	22
	Inconvénients	22
2.2.5	K-modes	22
	Principe de base	23
	Mesure de similarité	23
	Avantages	23
	Inconvénients	24
2.2.6	K-prototypes	24
	Principe de base	24
	Mesure de similarité entre individus	24
	Avantages	25
	Inconvénients	26
2.2.7	Classification ascendante hiérarchique	26
	Principe de base	26
	Mesure de similarité entre individus	27
	Stratégies d'agrégation : mesures de dissimilarité entre les clusters	27
	Avantages	29
	Inconvénients	29
3	Résultats	30
3.1	Analyse en clusters des consommateurs de cannabis ayant demandé un traitement pour la première fois au cours de l'année 2021 en Belgique	30
3.1.1	Analyse exploratoire des données	30
	Analyse des caractéristiques socio-démographiques des patients	30
	Analyse du profil d'addiction des patients	32
	Type de substances utilisées par les patients	34
3.1.2	Clustering	37
	K-means	37
	Classification ascendante hiérarchique	38
	K-modes	39
	K-prototypes	41
3.1.3	Comparaison des scores de silhouette issus des méthodes K-means, CAH, K-modes, K-prototypes et sélection de la meilleure méthode pour notre jeu de donnée	42
3.1.4	Interprétation des résultats	44
	Identification des caractéristiques clés des clusters	44
	Comparaison entre les clusters	48
3.2	Évolution des profils des patients au cours du temps : Comparaison des profils des patients entrés en traitement pour la première fois entre 2015 et 2020 avec ceux de l'année 2021	53
3.2.1	Évolution du nombre de patients enregistrés pour la première fois entre 2015 et 2021, pour un traitement lié à l'usage du cannabis	54
3.2.2	Évolution des profils des patients au cours du temps : comparaison des profils des patients entrés en traitement pour la première fois entre 2015 et 2020 avec ceux de l'année 2021	55
4	Discussion/Conclusion	62

Bibliographie	66
A Analyses complémentaires	68
B Code python	77

Remerciements

Il serait osé de prétendre remercier ici tous ceux qui ont contribué à la réalisation de ce mémoire de fin d'études. Aussi je considère la liste des personnes qui suit comme non exhaustive.

Je tiens en premier lieu à remercier ma promotrice Sciensano DR. ELENA DAMIAN, pour m'avoir accordé sa confiance et offert cette opportunité de réaliser ce travail au sein du département des drogues illégales chez Sciensano. Sa présence, ses nombreuses et pertinentes remarques ainsi que son soutien tant sur des aspects techniques qu'administratifs ont été d'un apport important pour ce travail.

J'adresse ensuite toute ma gratitude à JERÔME ANTOINE pour avoir mis sa grande expérience dans le traitement des données liées aux drogues illégales à mon service dans le cadre de ce mémoire. Sa disponibilité, ses nombreux conseils et remarques ainsi que sa relecture ont largement contribué à donner à ce travail sa consistance actuelle.

J'exprime également ma gratitude à ma promotrice UCLouvain PR. CATHERINE LEGRAND qui m'a accompagné tout au long de ce projet. Ses conseils avisés m'ont permis de donner à ce travail la bonne orientation et m'ont redonné le courage lorsque j'en suis venu à en manquer.

Je tiens enfin à remercier l'ensemble de mes amis du master en sciences des données, orientation statistiques de l'UCLouvain (Abouraman, Jean marc, Janice, etc...) avec qui j'ai passé de bons moments tout au long de mon master et qui m'ont apporté des remarques et des conseils lors de la rédaction de ce mémoire.

En dernier lieu mais certainement pas de la dernière importance, je remercie ici mon tendre époux CARLIN TCHUISSEU, pour sa présence et son soutien indéfectibles tout au long de mon parcours à l'UCLouvain. Sa compréhension lors des longues heures que j'ai consacrées dans la réalisation de ce mémoire a été un véritable soutien. Je suis reconnaissante d'avoir pu compter sur lui pour m'épauler, me motiver et me rappeler constamment la valeur de mes efforts.

Résumé

La consommation de cannabis en Belgique est assez répandue malgré le fait que la possession et la vente de cette substance restent illégales dans le pays. En effet, selon la dernière enquête HIS¹ réalisée en 2018, l'usage du cannabis est en augmentation dans la population belge ces dernières années : il a plus que doublé entre 2001 et 2018, passant de 10.7% à 22.6% [Gisle et Drieskens, 2019]. De plus, selon un rapport publié par sciensano en 2022, le cannabis fait partie des substances problématiques les plus fréquemment citées par les personnes usagères de drogues qui entrent en traitement dans toutes les régions [Antoine *et al.*, 2022]. La consommation de cannabis reste donc un sujet sensible qui nécessite une attention particulière de la part des autorités belges. De ce fait, connaître les différents profils des patients qui demandent un traitement pour une dépendance liée au cannabis peut être d'une importance capitale pour l'élaboration des programmes de prévention et des traitements ciblés contre la dépendance au cannabis.

Dans ce travail, nous identifions les clusters parmi les patients entrant en traitement pour la toute première fois sur la période 2015 à 2021, pour une dépendance au cannabis. Nos analyses se concentrent dans un premier temps sur l'année 2021. Ensuite, nous analysons la demande enregistrée sur la période 2015 à 2020 afin de voir si les profils de patients observés au cours de cette période sont similaires à ceux de l'année 2021. Dans le but d'identifier ces clusters, nous explorons différentes techniques de segmentation telles que le K-means, K-modes, K-prototypes et la Classification ascendante hiérarchique. Nous comparons les résultats de ces techniques et sélectionnons celle qui fournit le score de silhouette moyen le plus élevé, permettant ainsi d'obtenir le nombre optimal des clusters.

Les données proviennent de l'enregistrement TDI belge, qui tente de rassembler des informations lors de chaque épisode de traitement débuté par un patient dans un centre de traitement pour sa dépendance à l'alcool ou aux substances illicites.

Les variables de regroupement utilisées sont celles relatives aux indicateurs socio-démographiques des patients (âge, sexe, âge lors du premier usage du cannabis, niveau d'instruction, type de logement, etc...), à leur profil d'addiction (fréquence d'utilisation du cannabis etc...) et aux types de substances utilisées (alcool, cocaïne, autres stimulants, héroïne, hypnotiques, opiacés, etc...).

Deux clusters distincts ont été identifiés : Le premier comprend la majorité des patients (94.6%) et est principalement caractérisé par une dépendance aux substances moins

1. L'enquête HIS (Health Interview Survey) est une enquête nationale réalisée par Sciensano qui vise à collecter des données sur la santé et les comportements de la population belge, y compris des informations sur la consommation du cannabis. Elle est menée régulièrement, généralement tous les cinq ans et fournit des informations actualisées sur divers aspects de la santé des Belges. Il est généralement admis que cette enquête sous-estime la consommation de drogues, car elle se fait sur base volontaire et parce que la méthodologie utilisée ne permet pas d'inclure certains groupes de personnes particulièrement concernées par la consommation de drogues (personnes incarcérées, personnes sans domicile fixe etc...)

sévère (1 ou 2 substances problématiques). Le second cluster comporte les patients poly-problématiques (en ce sens qu'ils ont une dépendance à plusieurs substances et ont plus fréquemment eu recours à l'injection de substances). Nos analyses ont également montré une stabilité des profils des patients dans le temps. Les comparaisons des caractéristiques socio-démographiques, du profil d'addiction et du type de substances utilisées ont révélé des différences significatives entre ces clusters.

L'identification des clusters de patients demandant un traitement pour la première fois pour une dépendance au cannabis pourrait aider à personnaliser les traitements et les services qui leur sont offerts, en fonction de leurs besoins spécifiques. Cela pourrait également conduire à une amélioration de la qualité des soins et de l'efficacité des programmes de traitement liés à la dépendance au cannabis.

Abstract

The consumption of cannabis in Belgium is quite widespread despite the fact that the possession and sale of this substance remain illegal in the country. According to the latest Health Interview Survey (HIS)² conducted in 2018 by Sciensano, cannabis use has been increasing in the Belgian population in recent years, more than doubling from 10.7% in 2001 to 22.6% in 2018 [Gisle et Drieskens, 2019]. Additionally, a report published by Sciensano in 2022 indicates that cannabis is one of the most frequently cited problematic substances among drug users entering treatment across all regions [Antoine *et al.*, 2022]. The consumption of cannabis, therefore, remains a sensitive issue that requires special attention from the Belgian authorities. Consequently, understanding the different profiles of patients seeking treatment for cannabis addiction can be of paramount importance for the development of prevention programs and targeted treatments.

In this study, we aim to identify clusters among patients entering treatment for cannabis addiction for the first time between 2015 and 2021. Our analysis initially focuses on the year 2021, and then we examine the recorded demand over the period from 2015 to 2020 to determine if the patient profiles observed during this period are similar to those in 2021. To identify these clusters, we explore various segmentation techniques such as K-means, K-modes, K-prototypes, and hierarchical agglomerative clustering. We compare the results of these techniques and select the one that provides the highest average silhouette score, thus determining the optimal number of clusters.

The data used in this study are obtained from the Belgian Treatment Demand Indicator (TDI), which aims to collect information on each treatment episode initiated by a patient in a treatment center for alcohol or illicit substance dependence.

The clustering variables used include socio-demographic indicators of patients (age, gender, age at first cannabis use, education level, housing type, etc.), addiction profiles (cannabis use frequency, etc.), and types of substances used (alcohol, cocaine, other stimulants, heroin, hypnotics, opioids, etc.).

Two distinct clusters have been identified : the first cluster comprises the majority of patients (94.6%) and is primarily characterized by less severe substance dependence (1 or 2 problematic substances). The second cluster consists of polyproblematic patients (meaning they have dependence on multiple substances and more frequently engage in substance injection). Our analyses have also revealed stability in patient profiles over

2. The HIS (Health Interview Survey) is a national survey conducted by Sciensano aimed at collecting data on the health and behaviors of the Belgian population, including information on substance consumption such as cannabis. It is conducted regularly, usually every five years, and provides updated information on various aspects of the health of Belgians. It is generally accepted that this survey underestimates drug consumption because it is based on voluntary participation and because the methodology used does not allow for the inclusion of certain groups that are particularly affected by drug consumption (incarcerated individuals, homeless individuals, etc.)

time. Comparisons of socio-demographic characteristics, addiction profiles, and types of substances used have shown significant differences between these clusters.

The identification of clusters among patients seeking treatment for cannabis addiction for the first time could help personalize the treatments and services offered to them based on their specific needs. This could lead to improved quality of care and the effectiveness of cannabis addiction treatment programs.

Liste des figures

2.1	Représentation de la structure des données	10
2.2	Représentation de la distance Euclidienne	13
2.3	Représentation de la distance de Manhattan	14
2.4	Représentation de la distance de Hamming	15
2.5	Utilisation de la technique du coude pour choisir le nombre optimal de clusters	17
2.6	Illustration du calcul de l'indice de silhouette	18
2.7	Exemple de diagramme de silhouette	19
2.8	Illustration des étapes de l'algorithme K-means	21
2.9	Classification ascendante hiérarchique appliquée sur 5 points	27
2.10	Illustration du saut minimum	28
2.11	Illustration du saut maximum	28
3.1	Distribution de l'âge des patients et leur âge lors du premier usage du cannabis	31
3.2	Répartition de l'origine du renvoi en traitement et de la fréquence de consommation du cannabis chez les patients	33
3.3	Répartition du nombre total de substances problématiques par catégories	36
3.4	Analyse des tracés de silhouette (pour $k = 2$)	38
3.5	Analyse des tracés de silhouette (pour $k = 2$)	39
3.6	Répartition des âges (catégories)	40
3.7	Analyse des tracés de silhouette (pour $k = 3$)	41
3.8	Analyse des tracés de silhouette (pour $k = 2$)	42
3.9	Scores de silhouette moyen en fonction du nombre de clusters	43
3.10	Dendrogramme de la classification ascendante hiérarchique	44
3.11	Évolution du nombre de patients enregistrés pour la première fois entre 2015 et 2021, pour un traitement lié à l'usage du cannabis	54
3.12	Répartition des patients au sein des 3 clusters : année 2019	55
A.1	Analyse des tracés de silhouette(pour $k \in \{3, 4, 5, 6, 7\}$)	68
A.2	Analyse des tracés de silhouette(pour $k \in \{3, 4, 5, 6, 7\}$)	69
A.3	Elbow plot : kmodes	69
A.4	Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2015	72
A.5	Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2016	73

A.6	Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2017	74
A.7	Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2018	75
A.8	Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2020	76

Liste des tableaux

3.1	Répartition des patients en fonction du type de substances utilisées	34
3.2	Kmeans : score de silhouette moyen en fonction du nombre de clusters . . .	37
3.3	CAH : score de silhouette moyen en fonction du nombre de clusters	38
3.4	score de silhouette moyen en fonction du nombre de clusters : Kmodes . . .	40
3.5	score de silhouette moyen en fonction du nombre de clusters : Kprototypes	41
3.6	modalités des variables qui caractérisent au mieux le cluster 1	47
3.7	modalités des variables qui caractérisent au mieux le cluster 2	48
3.8	Différences entre les caractéristiques socio-démographiques des patients au sein des clusters	50
3.9	Différences entre les types de substances utilisées au sein des clusters . . .	52
3.10	Différences entre la fréquence d'utilisation du cannabis et l'origine du renvoi en traitement au sein des clusters	53
3.11	modalités des variables qui caractérisent au mieux le cluster 1	56
3.12	modalités des variables qui caractérisent au mieux le cluster 2	57
3.13	modalités des variables qui caractérisent au mieux le cluster 3	57
3.14	Différences entre les caractéristiques socio-démographiques des patients au sein des clusters : 2019	59
3.15	Différences entre les types de substances utilisées au sein des clusters : 2019	60
3.16	Différences entre la fréquence d'utilisation du cannabis et l'origine du renvoi en traitement au sein des clusters : 2019	61
A.1	Répartition des patients en fonction des caractéristiques sociodémographiques	70
A.2	Répartition des patients en fonction du profil d'addiction	71
A.3	Scores de silhouette moyen en fonction du nombre de clusters : année 2015	71
A.4	Score de silhouette moyen en fonction du nombre de clusters : année 2016 .	72
A.5	Score de silhouette moyen en fonction du nombre de clusters : année 2017 .	73
A.6	Score de silhouette moyen en fonction du nombre de clusters : année 2018 .	74
A.7	Score de silhouette moyen en fonction du nombre de clusters : année 2020 .	75

Liste des abréviations et définitions

Liste des abréviations

1. **EMCDDA** : European Monitoring Centre for Drugs and Drug Addiction
2. **FAMD** : Factorial analysis of mixed data
3. **TDI** : Treatment Demand Indicator

Définitions utilisées dans le contexte de ce travail :

1. **Centre de traitement** : Établissement ou praticien proposant un traitement pour les problèmes de consommation de substances illicites. Ce centre peut exercer en tant que service ambulatoire ou résidentiel, être spécialisé dans le traitement de l'addiction ou faire partie d'une gamme de services plus large visant différents groupes de patients, être médicalisé ou non. [[Antoine, 2022](#)]
2. **Patient** : Personne sans restriction d'âge, de nationalité ou de lieu de résidence ayant eu un contact direct avec un centre de traitement pour son problème de consommation de cannabis.
Les personnes ayant eu un contact uniquement par téléphone, lettre, Internet ou via leur famille sont exclues.
En raison de la protection de la vie privée, le patient doit être informé de l'enregistrement des données. Il doit également être informé de l'existence et des objectifs de l'enregistrement, des coordonnées de la personne responsable des données, de la destination des données, ainsi que de son droit d'accès et de rectification à ses données. Un patient peut renoncer à participer à l'enregistrement en le mentionnant par écrit.
3. **Traitement** : Toute activité visant à réduire ou éliminer les problèmes de consommation de cannabis ou d'autres substances chez une personne. Il inclut la désintoxication ou l'abstinence, le traitement de substitution, la pharmacothérapie, les programmes de réadaptation à long terme, la psychothérapie, le service de conseil, les traitements structurés avec une forte composante sociale, les traitements médicalement assistés, les interventions non médicales, le traitement spécifique en prison, ou les interventions visant à réduire les dommages liés aux substances si elles sont incluses dans un programme planifié.
Les interventions sporadiques qui ne font pas partie d'un programme planifié ne sont pas considérées comme faisant partie du traitement.

4. Types de substances :

Les substances prises en compte dans le TDI sont :

- La catégorie des opiacés, qui inclut l'héroïne, la méthadone détournée, la buprénorphine, le fentanyl illicite ou d'autres opioïdes.
- La catégorie de la cocaïne, qui inclut la cocaïne en poudre, le crack ou d'autres formes de cocaïne.
- La catégorie des stimulants (autres que la cocaïne), qui inclut les amphétamines, les méthamphétamines, la MDMA ou ses dérivés, la méphédronne ou d'autres stimulants.
- La catégorie des hypnotiques et des sédatifs qui inclut les barbituriques détournés, les benzodiazépines détournées, le GBH/GBL ou d'autres hypnotiques ou sédatifs détournés.
- La catégorie des hallucinogènes qui inclut le LSD, la kétamine ou d'autres hallucinogènes.
- Les inhalants volatils.
- La catégorie du cannabis incluant la marijuana (herbe), le haschich (résine) ou d'autres types de cannabis (ex. huile de haschich, cannabinoïdes de synthèse).
- L'alcool.
- La catégorie des autres substances non incluses dans les catégories ci-dessus.

Les patients inclus dans cette étude sont ceux dont le cannabis est la **substance principale** à l'origine de leur demande de traitement et qui débutent leur traitement pour la première fois. Ainsi, si un patient signale des problèmes liés à d'autres substances que le cannabis, celles-ci sont considérées comme des **substances problématiques**.

Introduction

La dépendance aux drogues est un grave problème de santé publique ; elle affecte toutes les sphères de la vie de la personne usagère de drogue mais se révèle également dangereuse pour son entourage. De plus, elle constitue un poids pour la société. Eu égard aux problèmes que représente ce trouble à l'échelle communautaire, l'Union européenne a mis sur pied l'observatoire européen des drogues et toxicomanies (EMCDDA) dans le but de centraliser l'information sur les drogues en Europe. Cette dernière a mis sur pied l'Indicateur de Demande de Traitement (TDI) afin d'évaluer le traitement de l'usage des drogues et des assuétudes en Europe. Le TDI fournit un format commun pour rapporter les données sur les personnes qui entreprennent un traitement en raison de leur consommation de drogue et de problèmes connexes au cours de chaque année civile. Depuis 2011, la Belgique collecte des données pour le TDI à travers le pays et celles-ci sont analysées par son institut de santé publique, Sciensano, qui publie régulièrement des études sur le sujet.

Par exemple, selon les résultats de la dernière enquête HIS réalisée en 2018, près de 23% de la population belge âgée de 15 à 64 ans a déjà consommé du cannabis au moins une fois, 7% au cours des 12 derniers mois, 4.3% au cours des 30 derniers jours et 3.1% l'utilisent de manière problématique (avec risques) [Gisle et Drieskens, 2019]. Environ 1.3% de la population belge présente un usage "intensif" du cannabis (consommation d'au moins 20 jours au cours des 30 derniers jours). De plus, un quart des jeunes de 15 à 24 ans ont déjà consommé le cannabis, tandis que la tranche d'âge des 24-34 ans présente la plus grande proportion de personnes ayant expérimenté cette substance (38.1%). En moyenne, la première consommation de cannabis a lieu vers l'âge de 19 ans.

Le rapport sur l'enregistrement TDI pour l'année 2021, publié par Sciensano, met également en évidence le cannabis en tant que substance problématique. Selon ce rapport, le cannabis est cité comme une substance problématique par 30% des personnes qui entrent en traitement, et comme substance principale par 15%. En Flandre, le cannabis est cité comme substance principale deux fois plus souvent (17%) qu'en Wallonie et à Bruxelles (8%) et ces proportions sont restées stables depuis 2015 [Antoine, 2022]. Ces études confirment que l'usage du cannabis est un problème de santé publique en Belgique. Par conséquent, l'analyse des profils des consommateurs de cannabis qui recherchent un traitement pour leur dépendance peut être bénéfique pour améliorer les approches d'intervention, de prévention et de traitement, afin de mieux répondre à leurs besoins particuliers.

Dans le cadre de ce mémoire de fin d'études réalisé en collaboration avec Sciensano, nous nous intéressons à la caractérisation des profils des personnes ayant sollicité un traitement pour leur dépendance au cannabis en Belgique entre 2015 et 2021 et n'ayant jamais suivi de traitement auparavant. Nous nous pencherons sur la question d'intérêt suivante : existe-t-il des clusters clairement identifiables parmi les consommateurs de cannabis ? pour chercher à identifier ces clusters, nous utiliserons des informations sur la (les)

substance(s) problématique(s) consommée(s), le profil d'addiction (ex : fréquence d'utilisation du cannabis, l'origine du traitement) et les caractéristiques sociodémographiques (ex : le sexe, l'âge, le niveau d'instruction, la situation du logement etc...).

Ce mémoire s'articule autour de 4 chapitres. Dans le premier, nous présentons les données ainsi que la méthodologie utilisées pour répondre à la question mentionnée plus haut. Le chapitre 2 présente un état de l'art des concepts théoriques abordés dans cette étude. Dans le chapitre 3, nous présentons les résultats obtenus. Nous nous concentrerons dans un premier temps sur les patients ayant sollicité un traitement pour la première fois pour leur addiction au cannabis au cours de l'année 2021. Ensuite, nous élargirons notre perspective en effectuant une analyse dans le temps : nous examinerons les profils des patients ayant demandé un traitement pour leur addiction au cannabis sur la période de 2015 à 2020, afin de déterminer si les clusters identifiés pour l'année 2021 persistent ou s'ils évoluent au fil du temps. Le dernier chapitre est dédié à la discussion et la conclusion de nos recherches.

Chapitre 1

Données et méthodologie utilisées dans cette étude

1.1 Données

1.1.1 Présentation des données

La base de données utilisée dans cette étude est extraite de l'enregistrement TDI belge. L'enregistrement TDI est un système de surveillance de la demande de traitement pour les troubles liés à l'utilisation de substances. Il a été développé par l'EMCDDA afin de fournir des informations sur les tendances de la demande de traitement dans les pays membres de l'Union Européenne. Ce registre rassemble des données transversales sur tous les épisodes de traitement débutés au cours d'une année donnée pour un problème de dépendance lié à l'alcool ou aux drogues illégales. Parmi les données recueillies auprès des patients, deux types d'informations sur les substances consommées sont collectées : tout d'abord, le patient est interrogé sur **toutes** les substances qui lui causent actuellement des problèmes. Ensuite, parmi ces substances, le patient est interrogé sur celle qui lui cause **le plus de problèmes** ou celle pour laquelle le traitement a été initié. Cette substance unique sera par la suite appelée « **substance principale** ». Si un patient a des problèmes avec plusieurs substances et ne peut identifier une seule substance principale, le champ correspondant à cette question dans le formulaire TDI reste vide [Antoine *et al.*, 2016].

Les données sur les patients ayant mentionné le cannabis comme substance principale au cours des années 2015 à 2021¹ ont été extraites ainsi que les variables qui correspondent aux caractéristiques socio-démographiques, au profil d'addiction et aux types de substances utilisées par les patients. Celles-ci sont présentées dans la section suivante.

1.1.2 Variables utilisées dans cette étude

Les variables utilisées dans cette étude sont réparties en trois groupes :

- **Les caractéristiques socio-démographiques :**

1. Durant cette période, **27669** personnes ont été enregistrées pour un traitement lié à l'usage du cannabis en Belgique (ce nombre inclut les patients ayant eu à demander un traitement auparavant).

CD_SEX : Le sexe

NR_AGE : L'âge du patient.

MS_AGE_FIRST_USE : l'âge du patient lors du premier usage du cannabis.

CD_LVN_STAT_WHERE : le type de logement (obtenue en posant la question : "Au cours des 30 derniers jours où résidiez-vous la plupart du temps ?" les catégories sont : "dans un domicile fixe", "dans des logements variables", "dans la rue", "en institution", "en prison", "dans un autre type d'endroit", "autre").

CD_LVN_STAT_WITH_WHOM : Le type de ménage (obtenue en posant la question "Au cours de ces 30 derniers jours, avec qui viviez-vous la plupart du temps ?" les catégories sont : "seul", "en couple", "avec un/mes parent(s)", "avec des autres membres de ma famille", "avec des amis ou autres personnes (sans lien de parenté)", "autre").

CD_LVN_CHILDREN : Variable dichotomique qui indique si le patient avait la responsabilité des enfants de moins de 18 ans au cours des 30 derniers jours ou non.

CD_EDUCATION_LEVEL : Le niveau d'instruction ("aucun", "primaire", "secondaire", "supérieur/université", "autre").

CD_LABOUR_STAT : La situation de travail ("emploi régulier", "emploi occasionnel", "au chômage", "écolier / étudiant / en formation", "incapacité de travail", "Homme / femme au foyer", "Pensionné / Pré-pensionné", "autre").

CD_INCOME_STAT : La source de revenus principale au cours des 30 derniers jours ("salaire / revenus du travail", "allocation de chômage", "pension de retraite ou de survie", "indemnité pour maladie ou invalidité", "revenu minimum ou support du CPAS", "allocations familiales (liées aux enfants)", "bourse d'études", "aucun revenu propre", "autre").

- **Les types de substances utilisées :**

Toutes les variables citées dans cette catégorie sont des variables dichotomiques.

Les variables *FL_OPIATES*, *FL_HEROIN*, *FL_BUPRENORPHINE*, *FL_METHADONE*, *FL_FENTANYL*, *FL_OTHER_OPIATES* indiquent si le patient a des problèmes avec les opiacés et ses différents types (héroïne, burprenophine, méthadone, fentanyl ou d'autres types d'opiacés).

Les variables *FL_COCAINE*, *FL_POWDER_COCAINE*, *FL_CRACK*, *FL_OTHER_COCAINE* permettent d'indiquer si le patient a des problèmes avec la cocaïne et les types de cocaïne (cocaïne en poudre, crack ou d'autres types de cocaïne).

Les variables *FL_STIMULANTS*, *FL_AMPHETAMINES*, *FL_METHAMPHETAMINES*, *FL_MDMA*, *FL_MEPHEDRONE*, *FL_OTHER_STIMULANTS* indiquent si le patient a des problèmes avec les stimulants (autres que la cocaïne) et les types de stimulants (amphétamines, méthamphétamines, MDMA, mephedrone, et d'autres types de stimulants).

FL_HYPNOTICS, *FL_BARBITURATES*, *FL_BENZODIAZEPINES*, *FL_GHB*, *FL_OTHER_HYPNOTICS* correspondent aux problèmes avec les hypnotiques et les types d'hypnotiques.

FL_HALLUCINOGENS, *FL_LSD*, *FL_KETAMINE*, *FL_OTHER_HALLUCINOGENS* problèmes avec les hallucinogènes.

Les variables *FL_VOLATILE*, *FL_ALCOHOL*, *FL_OTHER_SUBSTANCE* permettent d'identifier si le patient a des problèmes respectivement avec les inhalants volatils, l'alcool ou d'autres types de substances.

A partir des substances problématiques mentionnées par le patient, nous avons créé une variable catégorielle nommée ***nombre_subst_total*** qui indique le nombre total de substances problématiques ("1-2", "3-5", "6-10"). Nous l'avons créé parce qu'elle nous sera utile pour mieux comprendre les caractéristiques des clusters identifiés. Elle est donc une variable supplémentaire et ne fait pas partie des variables utilisées pour identifier les clusters.

- **Le profil d'addiction :**

CD_FREQ_USE : la fréquence d'utilisation du cannabis le mois précédent ("Je ne l'ai pas consommée le mois précédent", "1 jour par semaine ou moins", "2 à 3 jours par semaine", "4 à 6 jours par semaine", "Tous les jours").

CD_INJECTING_STAT : cette variable indique si le patient a déjà consommé des substances psychoactives par injection au cours de sa vie (quelle que soit la substance). De même que la variable *nombre_subst_total* mentionnée précédemment, elle ne fait pas partie des variables utilisées pour identifier les clusters et sert juste à identifier les caractéristiques des clusters. Généralement les utilisateurs de cannabis ont tendance à le fumer/ inhaler . Cependant, le choix de prendre cette variable comme supplémentaire permettrait de caractériser les clusters de patients qui ont déjà eu recours aux substances injectables.

CD_REFERRAL : canal considéré comme déterminant par lequel le patient est entré en traitement ("moi-même", "quelqu'un de ma famille", "un ami", "un médecin généraliste", "un centre pour toxicomanes", "un hôpital", "un autre service médical ou psychosocial", "la justice", "autre").

En utilisant ces variables, nous avons identifié et caractérisé les clusters. La méthodologie utilisée est décrite dans la suite.

1.2 Méthodologie utilisée dans cette étude

1.2.1 Méthodes utilisées pour identifier les différents clusters

Il est important de remarquer que toutes les variables dont nous disposons sont des variables catégorielles, à l'exception de l'âge des patients et l'âge lors du premier usage du cannabis qui elles, sont continues.

Lorsqu'on veut faire du clustering, une approche courante consiste à d'utiliser les méthodes factorielles telles que l'analyse des Correspondances Multiples (lorsque les variables sont toutes catégorielles) et l'analyse factorielle des données mixtes ou AFDM (lorsque les données contiennent à la fois les variables continues et catégorielles) [Yong *et al.*, 2013].

Dans le cadre de notre étude, puisque nous disposons des données mixtes, nous avons donc dans un premier temps utilisé l'AFDM afin de transformer les données mixtes en un ensemble de caractéristiques continues (axes factoriels).

Détermination du nombre de facteurs à retenir suite à l'AFDM :

L'AFDM est généralement utilisée pour transformer les données en un espace dimensionnel inférieur tout en conservant une certaine quantité d'informations dans les données d'origine [Yong *et al.*, 2013]. Pour cela, en fonction des données, on peut retenir le nombre d'axes factoriels nécessaires pour capter le pourcentage de la variance des données que l'on souhaite conserver [Huang, 1998]. Ainsi, en retenant le nombre d'axes factoriels qui permet de capter 100% de la variance, on aura capté la totalité de l'information contenue dans les données.

Dans cette étude, puisque nous ne sommes pas intéressés par la conversion de nos données dans un espace dimensionnel inférieur, l'AFDM a été utilisée plutôt pour homogénéiser notre ensemble de données mixtes en convertissant toutes nos variables en valeurs numériques normalisées (c'est-à-dire que nous avons pris le nombre de variables synthétiques nécessaire pour capter 100% de la variabilité des données) [Ahmad et Khan, 2019]. L'AFDM nous a ainsi permis d'obtenir de nouvelles données dont les variables sont formées des axes factoriels et qui capturent la totalité de l'information contenue dans les données de départ.

Identification des différents clusters de patients :

Lorsqu'il s'agit de clustering, il n'y a pas de méthode standard qui garantisse des résultats optimaux dans tous les cas. En raison de la diversité des données et des structures sous-jacentes, il est donc généralement préférable de considérer plusieurs approches et de tester différentes méthodes afin de choisir celle qui produit les résultats les plus satisfaisants pour l'ensemble de données. De ce fait, dans le but d'identifier les différents clusters de patients, nous avons utilisé 3 approches différentes :

1. **Une approche qui utilise uniquement des données continues :** Dans ce cadre, nous avons implémenté les méthodes **Kmeans** et **CAH** sur les axes factoriels issus de l'AFDM [Lebart *et al.*, 1995, Pagès, 2002].

2. **Une approche qui utilise uniquement des données catégorielles : Kmodes**
Puisque cette méthode prend en compte uniquement les variables catégorielles, nous avons utilisé la base de donnée de départ et regroupé par catégories les variables continues (l'âge des patients et leur âge lors du premier usage du cannabis) afin de l'implémenter [Huang, 1998].
3. **Une approche basée sur les données mixtes : Kprototype**
Aucune transformation n'a été nécessaire sur les données de départ car elle prend directement en compte les données mixtes [Huang, 1998].

Chacune de ces méthodes présente des avantages et des inconvénients qui seront discutés dans le chapitre 2.

Les résultats fournis par les méthodes citées ci-dessus sont évalués et comparés à l'aide du **score de silhouette** [Zhou et Gao, 2014, Dinh *et al.*, 2019]. La méthode pour laquelle le score de silhouette moyen est le plus élevé est celle qui fournit la meilleure partition. Dans la littérature, il est courant de comparer les performances de différentes techniques de clustering en utilisant le score de silhouette. Bien que les distances utilisées dans les méthodes de clustering citées précédemment puissent différer (par exemple la distance Euclidienne pour K-means, la distance de Hamming pour K-modes etc...), le score de silhouette reste une mesure relative qui compare les partitions intra-cluster et inter-cluster pour évaluer la qualité générale du clustering. Il permet de quantifier la structure des données et la séparation des groupes, indépendamment de la distance utilisée (voir la section 2.2.3 pour plus de détails sur cette métrique).

Plusieurs auteurs ont utilisé le score de silhouette pour comparer les performances des techniques de clustering dans différentes études. Par exemple, Palani et al. [Palani *et al.*, 2021] ont comparé les performances de la classification ascendante hiérarchique, K-prototypes et les modèles de mélange Gaussien en utilisant le score de silhouette.

1.2.2 Méthodes utilisées pour caractériser les clusters identifiés

La caractérisation des clusters s'est faite en deux étapes :

conformément à la démarche souvent employée dans la littérature, la caractérisation des clusters s'est faite à l'aide du test exact de Fisher (en ce qui concerne les caractéristiques liées aux indicateurs socio-démographiques, aux types de substances utilisées et au profil d'addiction).

les différences de proportions entre les clusters sont ensuite évaluées à l'aide des tests du chi carré et du test de student (en ce qui concerne les caractéristiques liées aux âges) [Schönnesson *et al.*, 2008]. Le niveau de significativité considéré est de 5%.

1.2.3 Méthodes utilisées pour la comparaison des profils observés sur la période 2015 à 2020 à ceux observés sur l'année 2021

Dans le but d'identifier les clusters, nous avons extrait les données de chaque année et appliqué uniquement le méthode qui fournit le meilleur résultat sur l'année 2021. Cette approche permet de garder une cohérence dans les comparaisons. Le score de silhouette est également utilisé pour déterminer le nombre optimal de clusters.

Maintenant que nous avons présenté la base de données et la méthodologie utilisées dans notre étude, nous allons explorer dans le chapitre suivant les concepts théoriques relatifs au clustering.

Chapitre 2

Concepts théoriques et état de l'art

Lorsqu'on veut effectuer du clustering, il est souvent utile d'effectuer dans un premier temps une analyse factorielle sur des données, ce qui peut rendre les analyses plus rapides [Lebart *et al.*, 1995]. L'analyse factorielle vise à réduire les dimensions d'un tableau de données en représentant les associations entre individus et entre variables dans des espaces de faibles dimensions. Ainsi, en projetant l'ensemble des observations sur des axes factoriels de manière à limiter la perte d'information, l'analyse factorielle permet de mettre en avant des individus au comportement homogène relativement à une ou à plusieurs variables.

Dans notre étude l'utilisation de l'analyse factorielle sur les données a pour objectif de construire un nouveau système de représentation (axes factoriels) qui permet de capter toute l'information contenue dans les données, afin d'appliquer les méthodes classiques de clustering telles que le K-means et la classification ascendante hiérarchique.

Lorsqu'on veut effectuer une analyse factorielle, les techniques populaires telles que l'analyse en composantes principales (ACP) ou l'analyse des correspondances multiples (ACM), largement présentées dans la littérature sont généralement utilisées :

- L'ACP est utilisée lorsque toutes les variables actives sont quantitatives [Lebart *et al.*, 1995, Palm, 1998, Duby et Robin, 2006].
- L'ACM quant à elle est utilisée lorsque les variables sont toutes catégorielles [Lebart *et al.*, 1995, Palm, 2007].

Lorsque les données contiennent à la fois les variables continues et catégorielles, la situation est un peu plus complexe.

Une première approche consiste à utiliser les variables continues comme variables actives et utiliser une ACP, puis utiliser les variables qualitatives comme variables illustratives. Ou à l'inverse, utiliser les variables qualitatives comme variables actives et utiliser une ACM, puis utiliser les variables quantitatives comme variables illustratives. Cependant, les résultats obtenus via cette approche sont souvent critiqués dans la littérature car seulement une partie de l'information participe à la construction des facteurs [Lebart *et al.*, 1995].

Une seconde approche consiste à découper en classes les variables quantitatives et d'effectuer une ACM. Seulement, le découpage des classes passe par le choix du nombre d'in-

tervalles. Pour cela, il faut utiliser un algorithme afin de choisir les bornes de discrétisation. Ce processus est susceptible de dénaturer l'information traitée [Cadario et Parguel, 2014]. Dans la pratique, une approche consiste souvent à utiliser le contexte de l'étude dans le but de choisir les bornes des intervalles.

Une troisième approche consiste à remplacer les variables catégorielles par une série de variables indicatrices via un codage disjonctif complet. En procédant ainsi, l'information n'est pas dénaturée car il y a une bijection exacte entre les variables initiales et les variables recodées. L'analyse factorielle des données mixtes (AFDM) suit cette dernière approche [Pagès, 2002].

Dans ce chapitre, la section 1 présente le principe de l'analyse factorielle des données mixtes. Les concepts théoriques utilisés pour identifier et caractériser les différents clusters sont présentés dans la section 2.

2.1 Analyse factorielle des données mixtes (AFDM)

2.1.1 Principe de l'AFDM

La figure 2.1 ci-dessous tirée de [Lebart *et al.*, 1995] permet de visualiser la structure d'un tableau de données comportant n observations et p variables. On suppose que les variables 1 à $j - 1$ sont continues et les variables j à p sont catégorielles.

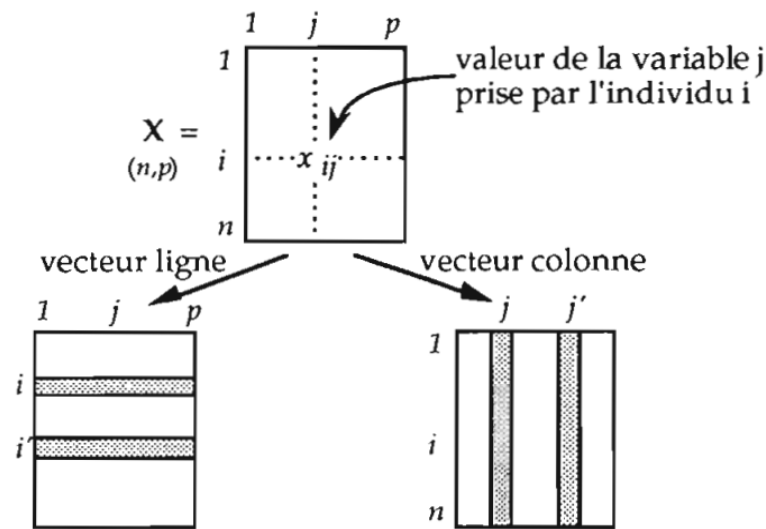


FIGURE 2.1 – Représentation de la structure des données

source : [Lebart *et al.*, 1995]

Pour construire un nouveau système de représentation qui préserve au mieux l'information contenue dans ce tableau de données, l'AFDM passe par l'application d'une analyse en composante principale sur des données transformées judicieusement. Les étapes de l'AFDM sont les suivantes :

- **Codage disjonctif complet des variables catégorielles :** Les variables quantitatives (1 à $j - 1$) sont laissées en l'état, les catégorielles (j à p) permettent d'obtenir l nouvelles variables indicatrices, où l est le nombre total de catégories contenues dans les variables j à p . Ces nouvelles variables sont combinées aux variables quantitatives (1 à $j - 1$), pour former un nouveau tableau de données dans lequel les variables 1 à $j - 1$ sont les variables continues initiales et les variables j à $j + l$ sont les variables indicatrices. Ce nouveau tableau est utilisé pour l'étape suivante.
- **Standardisation différenciée des colonnes :** Afin d'équilibrer l'influence de chaque type de variables, les variables quantitatives et indicatrices sont standardisées différemment :
 - Les variables continues sont centrées et réduites. Si Z_{ik} désigne la valeur prise par l'individu i , $i = 1, \dots, n$ pour la variable k , $k = 1, \dots, j - 1$, on a :

$$Z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k}; k = 1, \dots, j - 1$$

μ_k : moyenne de la variable k

σ_k : écart-type de la variable k .

- Les variables indicatrices sont normalisées par la racine carrée de la proportion de valeurs correspondant à "1" dans la colonne.
Par exemple, Si Z_{ik} désigne la valeur prise par l'individu i , $i = 1, \dots, n$ pour la variable k , $k = j, \dots, j + l$, on a :

$$Z_{ik} = \frac{x_{ik}}{\sqrt{p_k}}; k = j, \dots, j + l$$

$p_k = \frac{n_k}{n}$ et n_k , le nombre total de 1 dans la variable indicatrice k ; $k = j, \dots, j + l$

- **Utilisation de l'ACP sur le tableau de données résultant afin d'obtenir les axes factoriels.**

2.1.2 Intérêt de l'AFDM comme étape préalable à la classification

Pourquoi combiner l'AFDM et la classification : Dans la situation où on dispose d'un jeu de données contenant des variables mixtes, l'AFDM peut être utilisée pour transformer les variables initiales en variables continues (axes principaux) contenant les informations les plus importantes dans le jeu de données.

Comment combiner l'AFDM et la classification :

1. Effectuer l'AFDM sur les données en suivant les étapes définies précédemment, tout en choisissant le nombre de dimensions à retenir. Le nombre de dimensions à retenir dépend que la quantité d'information contenue dans le jeu de données que l'on veut préserver.
2. Appliquer les techniques de clustering que l'on souhaite sur le résultat de l'étape 1.

3. Interpréter les clusters à l'aide des variables initiales du jeu de données.

Le premier plan factoriel est généralement utilisé dans un but de visualisation des clusters formés. Toutefois, la caractérisation et l'interprétation des clusters se fait à l'aide des variables initiales.

Dans la suite, nous présentons quelques méthodes généralement utilisées pour effectuer le clustering.

2.2 Quelques méthodes de clustering

Le clustering est une technique d'apprentissage automatique non supervisé qui regroupe les instances en clusters en fonction de leurs similitudes. Par exemple, dans le cadre de notre étude, on veut identifier les groupes de patients ayant des caractéristiques similaires. Pour cela, on veut que :

- Les patients appartenant à un même groupe soient similaires (se ressemblent le plus possible)
- Les patients appartenant à deux groupes différents se démarquent le plus possible

Il est difficile d'apprécier le regroupement sans bien comprendre ce que cela signifie pour les patients d'être similaires. La section suivante présente quelques mesures de similarité couramment utilisées dans le clustering.

2.2.1 Mesures de similarité

Le choix d'une mesure de similarité pour effectuer le clustering est crucial car cela affecte directement la qualité des groupes obtenus [Fu *et al.*, 2004]. En effet, Si la mesure de similarité utilisée n'est pas adaptée à la structure des données, les groupes obtenus peuvent être imprécis ou incohérents. Il existe plusieurs mesures de similarités généralement utilisées dans le clustering :

La distance Euclidienne : La distance euclidienne est l'une des mesures de distance les plus courantes pour les données numériques. Elle calcule la distance entre deux points dans un espace euclidien en prenant la racine carrée de la somme des carrés des différences entre les coordonnées des deux points. La formule de la distance euclidienne entre deux points A et B dans un espace n-dimensionnel est la suivante :

$$d(A, B) = \sqrt{\sum_{i=1}^n (A_i - B_i)^2}$$

La figure 2.2 ci-dessous tirée de [Malik et Tuckfield, 2019] permet de visualiser le calcul de distance euclidienne dans le cas d'un espace à trois dimensions :

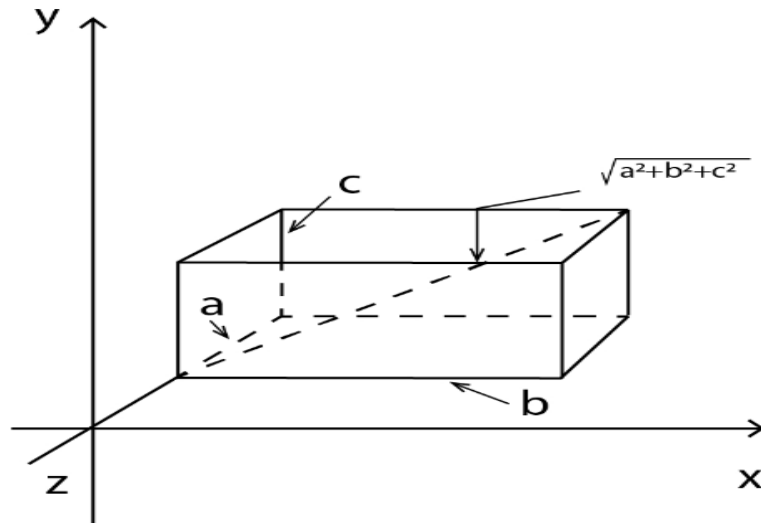


FIGURE 2.2 – Représentation de la distance Euclidienne

source :[Malik et Tuckfield, 2019]

La distance Euclidienne présente des avantages mais aussi des inconvénients :

Avantages : Cette distance est simple et facile à calculer, interprétable dans l'espace euclidien. **Inconvénients :** La distance Euclidienne est sensible aux valeurs aberrantes (outliers) et à l'échelle des données. Elle est Non-applicable aux données catégorielles ou discrètes. Une façon de palier à ces limites est d'utiliser la distance de Manhattan.

La distance de Manhattan : La distance de Manhattan est une mesure de distance qui calcule la distance entre deux points en prenant la somme des différences absolues de leurs coordonnées. Elle ne majore donc pas la pondération des outliers. En revanche, les temps de calcul sont particulièrement longs. La formule de la distance de Manhattan entre deux points A et B est la suivante :

$$d(A, B) = |A_1 - B_1| + |A_2 - B_2|$$

La figure 2.3 ci-dessous tirée de [Malik et Tuckfield, 2019] permet de visualiser de calcul de distance euclidienne dans le cas d'un espace bidimensionnel :

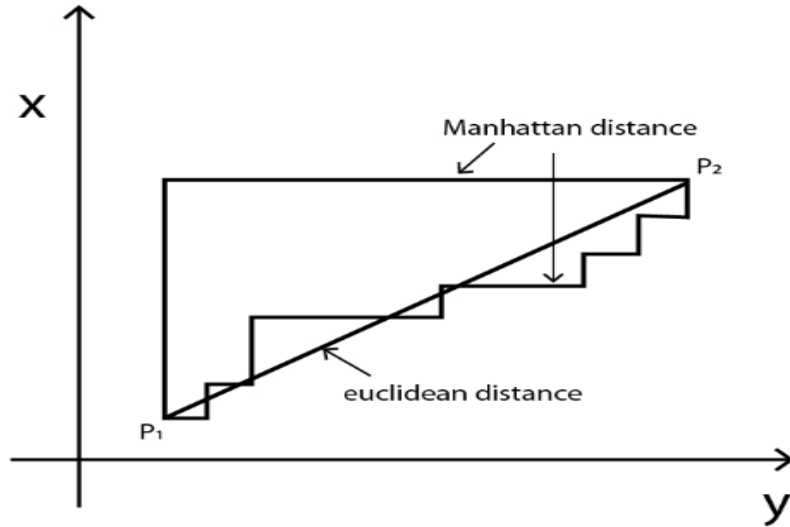


FIGURE 2.3 – Représentation de la distance de Manhattan

source :[[Malik et Tuckfield, 2019](#)]

La longueur de la ligne diagonale est la distance euclidienne entre les deux points. La distance de Manhattan est simplement la somme de la valeur absolue des différences entre deux coordonnées. Ainsi, la principale différence entre la distance euclidienne et la distance de Manhattan est qu'avec la distance euclidienne, nous mettons au carré les distances entre les coordonnées, puis prenons la racine de la somme, mais à la distance de Manhattan, nous prenons directement la somme de la valeur absolue des différences entre coordonnées.

Les distances définies précédemment sont généralement utilisées lorsque les données sont continues, Cependant, bon nombre de bases de données contiennent à la fois des données catégorielles ou binaires. Dans ce cas, une solution est d'utiliser la distance de Hamming.

La distance de Hamming : La distance de Hamming est une mesure de distance souvent utilisée pour les données binaires ou catégorielles. Elle calcule la distance entre deux chaînes de caractères en comptant le nombre de positions où les caractères sont différents. La formule de la distance de Hamming est la suivante :

$$d(A, B) = \sum_{i=1}^n i, \quad \text{Si } A_i \neq B_i$$

En d'autres termes, étant donné deux points de dimensions égales, la distance de Hamming est le nombre de coordonnées différentes l'une de l'autre. La figure 2.4 ci-dessous tirée de [[Malik et Tuckfield, 2019](#)] illustre le calcul de la distance de hamming dans le cas des données binaires.

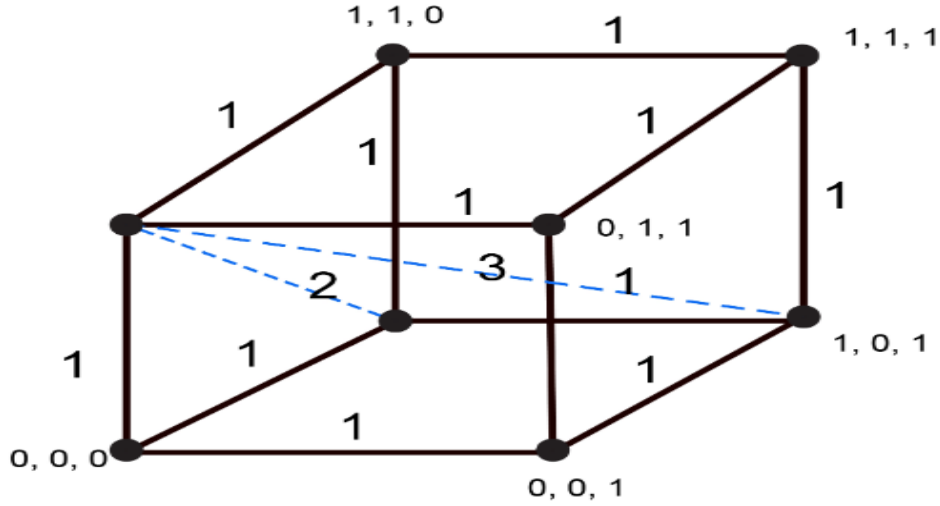


FIGURE 2.4 – Représentation de la distance de Hamming

source :[[Malik et Tuckfield, 2019](#)]

Si on prend deux points $(0, 1, 1)$ et $(0, 1, 0)$. Une seule valeur (la dernière valeur) est différente entre ces deux variables. Ainsi, la distance de Hamming est de 1.

Tout comme la distance Euclidienne et de Manhattan, elle présente des avantages et des inconvénients :

Avantages : Cette distance est appropriée pour les données binaires, où chaque dimension ne peut prendre que deux valeurs (0 ou 1) et utile pour la comparaison de chaînes de caractères.

Inconvénients : Elle ne tient pas compte des corrélations entre les dimensions et est Non-applicable aux données continues.

En plus de s'assurer que les patients appartenant à chaque cluster soient similaires, il faut également tenir compte de l'inertie intra-cluster et de l'inertie inter-cluster. Ces deux notions sont définies dans la suite.

2.2.2 inertie intra-cluster et inter-cluster

L'inertie est une mesure de la quantité de variance expliquée par un modèle de clustering. **L'inertie intra-cluster** mesure la dispersion des données à l'intérieur d'un cluster, tandis que **l'inertie inter-cluster** mesure la dispersion entre les clusters.

L'inertie totale d'un nuage $\Gamma = \{M_i, \quad i = 1, \dots, n\}$ est la somme pondérée des carrés des distances de ses points au centre de gravité du nuage [[Lebart et al., 1995](#)]. Donc, si G désigne le centre de gravité de Γ , l'inertie totale de Γ est, si tous les points du nuage sont de même poids égal à $\frac{1}{n}$, donnée par la formule :

$$I_{totale} = I(\Gamma) = \left(\frac{1}{n}\right)(d(M_1, G)^2 + d(M_2, G)^2 + \dots + d(M_n, G)^2) \quad (2.1)$$

où d désigne une distance choisie.

Lorsque le nuage est très dispersé, l'inertie totale est grande. Par contre, si le nuage est

constitué de points bien regroupés, celle-ci est petite.

Si le nuage Γ est composé de k clusters $\Gamma_1, \Gamma_2, \dots, \Gamma_k$ (disjointes deux à deux), celles-ci seront d'autant plus homogènes que les inerties de chaque cluster $I(\Gamma_1), I(\Gamma_2), \dots, I(\Gamma_k)$, calculées par rapport à leurs centres de gravité $G_1, G_2, \dots, G_k$ respectifs, seront faibles. L'inertie intra-cluster sera donc la somme de ces inerties :

$$I_{intra} = I(\Gamma_1) + I(\Gamma_2) + \dots + I(\Gamma_k)$$

Dans cette relation, les inerties $I(\Gamma_1), I(\Gamma_2), \dots, I(\Gamma_k)$ sont calculées à l'aide de la formule 2.1, dans laquelle le centre de gravité G du nuage est remplacé par $G_1, G_2, \dots, G_k$ et le poids $\frac{1}{n}$ par celui du cluster.

L'inertie inter-cluster mesure la dispersion des clusters par rapport au centre de gravité du nuage :

$$I_{inter} = p_1 d(G_1, G)^2 + p_2 d(G_2, G)^2 + \dots + p_n d(G_n, G)^2$$

où $p_j = \frac{1}{n_j}$ et n_j , l'effectif du cluster j ($j = 1, \dots, n$).

L'inertie totale du nuage de points $I(\Gamma)$ est la somme de son inertie inter-cluster et intra-cluster :

$$I(\Gamma) = I_{intra} + I_{inter}$$

Comme mentionné précédemment, la partition d'un nuage en k clusters sera d'autant meilleure que son inertie intra-clusters sera petite, ou, puisque son inertie totale reste la même quelque soit la partition, que son inertie inter-clusters est grande.

Un bonne classification est celle pour laquelle il y'a une :

- petite variabilité intra-cluster (c'est-à-dire une petite distance entre les individus d'un même groupe)
- grande variabilité inter-cluster (c'est-à-dire une grande distance entre les groupes différents)

La détermination du nombre optimal de clusters consiste donc à trouver le nombre de groupes qui satisfait au mieux à ces deux critères. Quelques méthodes couramment utilisées pour le choix du nombre de clusters sont présentées dans la suite.

2.2.3 Choix du nombre optimal de clusters

Trouver le nombre de clusters correcte reste un problème fondamental des méthodes de clustering. Pour ce faire, il existe plusieurs méthodes mais une des approches les plus couramment utilisées est basée sur les mesures de similarité. Cette méthode consiste à utiliser des mesures de qualité du clustering, comme la technique du coude ou l'indice de silhouette, afin d'évaluer la qualité des clusters pour différents nombres de clusters. Ces deux techniques sont présentées dans la suite.

Technique du coude

La technique du coude [Humaira et Rasyidah, 2020] consiste à tracer la relation entre le nombre de clusters et la somme des carrés des distances intra-cluster. Elle utilise l'hétérogénéité au sein de chaque cluster pour évaluer la variabilité au sein de celui-ci. En faisant varier le nombre de clusters, on s'intéresse au pourcentage de la variance expliquée par chaque groupe. On s'attend à ce que l'hétérogénéité diminue. Le défi est de trouver la valeur de k qui dépasse les rendements décroissants. Une fois cette valeur atteinte, l'ajout d'un nouveau cluster n'améliore pas la variabilité des données car il reste très peu d'informations à expliquer.

Dans la pratique, cette méthode consiste à effectuer une série d'applications (K-means, K-modes, etc...) en faisant largement varier k .

Pour chaque valeur de k , on récupère les valeurs de l'inertie intra-cluster puis on trace celles-ci en fonction de k (on peut aussi travailler sur la variabilité inter-cluster et les tracer en fonction de k). La figure 2.5 ci-dessous tirée de [Coquillard, 2018] permet de visualiser un exemple.

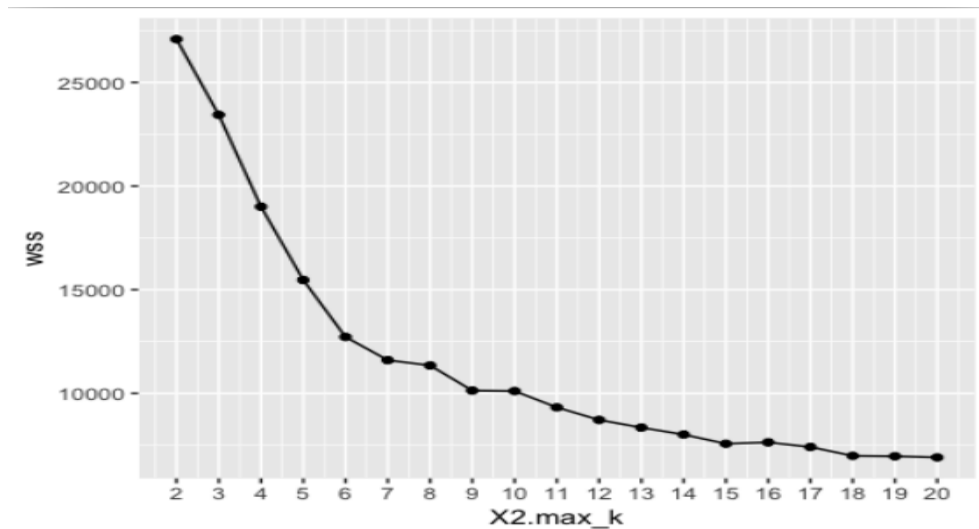


FIGURE 2.5 – Utilisation de la technique du coude pour choisir le nombre optimal de clusters

source : [Coquillard, 2018]

Dans cet exemple, la valeur de k optimale est vraisemblablement 7 car au delà, la variabilité diminue beaucoup plus lentement.

Une autre façon d'analyser la qualité de la partition est d'utiliser le score de silhouette.

Score de silhouette

Le score de silhouette (ou indice de silhouette) mesure à quel point les individus d'un même cluster sont similaires. Plus le score de silhouette est proche de 1, meilleure est la qualité du cluster. En utilisant le score de silhouette pour différents nombres de clusters, on peut déterminer le nombre optimal de clusters, qui maximise la qualité des clusters. C'est un excellent outil pour inspecter visuellement les similitudes au sein des clusters et

les différences entre les clusters [Zhou et Gao, 2014]. La figure 2.6 ci-dessous montre un exemple du calcul de l'indice de silhouette.

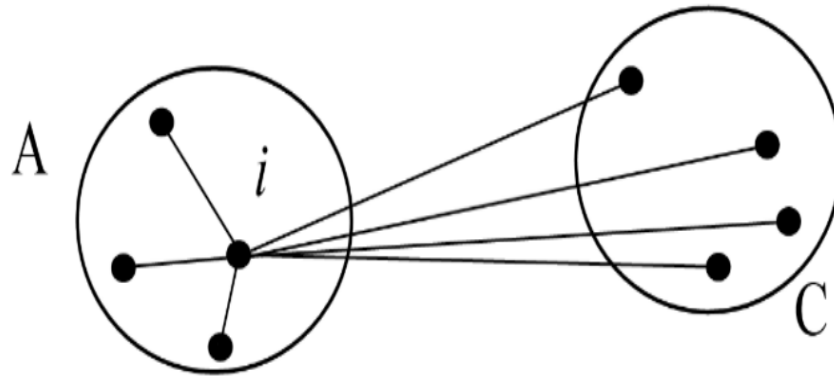


FIGURE 2.6 – Illustration du calcul de l'indice de silhouette

source : [Mamat et al., 2018]

La figure 2.6 présente deux clusters notés A et C. Chacun d'eux contient des points de données. En prenant par exemple le cluster A, on considère un point i appartenant à ce cluster et on calcule son indice de silhouette par la formule :

$$S(i) = \frac{(b(i) - a(i))}{\max(a(i), b(i))}$$

où :

- $S(i)$ est l'indice de silhouette du point i
- $a(i)$ est la distance moyenne entre i et tous les autres points du cluster auquel i appartient (le cluster A dans ce cas).
- $b(i)$ est la distance moyenne minimale entre i et tous les clusters auxquels i n'appartient pas (le cluster C dans ce cas).

$S(i)$ est compris entre -1 et 1.

- Si la valeur de la silhouette $S(i)$ est proche de 1, l'échantillon est bien regroupé et déjà affecté à un cluster très approprié.
- Si $S(i)$ est nul, l'échantillon peut être attribué à un autre cluster le plus proche et l'échantillon se trouve à égale distance des deux clusters. Cela signifie qu'il indique des clusters qui se chevauchent.
- Si $S(i)$ est proche de -1, l'échantillon est mal classé et est simplement placé quelque part entre les clusters.

Le score de silhouette est calculé pour chaque observation des clusters A et C et **la moyenne de tous les indices de silhouette** est égale au score de silhouette global de la partition.

Lors du calcul du score de silhouette moyen, les valeurs négatives sont prises en compte de la même manière que les valeurs positives. Le score de silhouette moyen est calculé en prenant la moyenne de tous les scores de silhouette individuels des observations présentes dans le jeu de données [Rousseeuw, 1987]. Ainsi, lorsque des observations ont des scores de silhouette négatifs, ceux-ci réduisent la valeur globale du score de silhouette moyen, car ils contribuent à la somme totale des scores de silhouette.

Les diagrammes de silhouette typiques représentent l'étiquette de cluster sur l'axe y et le score de silhouette sur l'axe des x. La taille/épaisseur des silhouettes est également proportionnelle au nombre d'observations à l'intérieur du cluster. La figure 2.7 ci-dessous permet de visualiser le diagramme de silhouette d'une partition en 3 clusters.

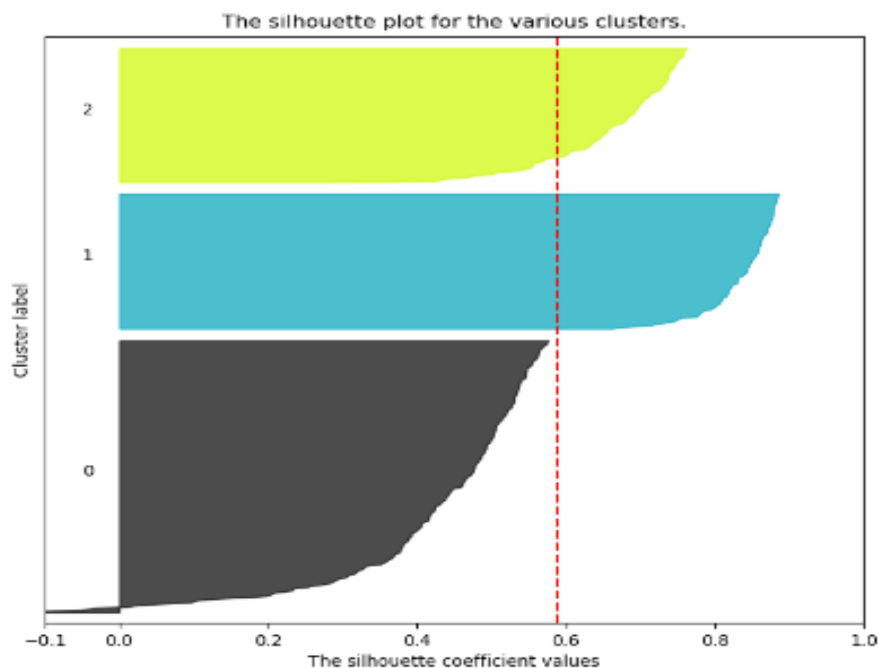


FIGURE 2.7 – Exemple de diagramme de silhouette

On peut observer sur la figure 2.7 que certains points appartenant au plus grand cluster (étiqueté '0') sont mal classés car leur indice de silhouette est négatif. De plus, le score de silhouette moyen se situe autour de 0.58, ce qui dénote une bonne partition.

Dans la suite nous présentons différentes techniques de clustering. Nous verrons trois méthodes basées sur les centroïdes (la méthode des k-moyennes, des K-modes, et des K-prototypes) et une méthode basée sur la hiérarchie (la classification ascendante hiérarchique).

2.2.4 K-means

L'algorithme des k-moyennes(K-means) s'applique à des situations où les variables sont numériques(et habituellement standardisées). L'objectif de cette méthode est de par-

tionner les données en k groupes où k est une valeur fixée à l'avance. En général, le nombre de clusters contenus dans les données n'est pas connu à priori.

Principe de base

. Cet algorithme est relativement simple à mettre en oeuvre et son principe est qu'à chaque étape de son exécution, l'inertie intra-cluster est diminuée (voir la partie 2.2.2 pour la définition de l'inertie intra-cluster).

Les étapes de cet algorithme sont les suivantes :

1. Initialisation : Déterminer le nombre de clusters k et sélectionner k centres initiaux de manière aléatoire.
2. Assignation : Pour chaque point de données, calculer la distance entre le point et chaque centre du cluster. Affecter le point au cluster dont le centre est le plus proche (généralement à l'aide de la distance euclidienne).
3. Mise à jour : Calculer le nouveau centre pour chaque cluster en prenant la moyenne de tous les points qui y sont affectés.

Les coordonnées des centroïdes (le vecteur-moyenne) pour chacun des k clusters, sont données par la formule :

$$\mu_k = \frac{1}{N_k} \sum_{i:C(i)=k} x_i, \quad k = 1, \dots, k$$

où N_k est le nombre d'observations dans le cluster k .

4. Répéter les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de changement dans les affectations de clusters ou qu'un nombre maximum d'itérations soit atteint.

Ces étapes sont illustrées par la graphe suivant :

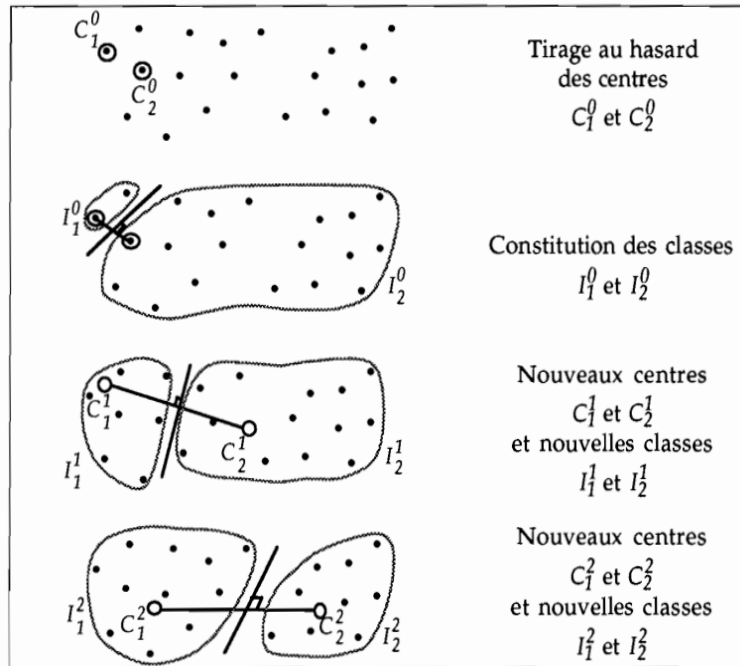


FIGURE 2.8 – Illustration des étapes de l'algorithme K-means

source : [Lebart *et al.*, 1995]

Sur la figure 2.8, l'algorithme commence par choisir deux centres de clusters provisoires, de manière aléatoire (C_1^0 et C_2^0). Cela induit une première partition des individus en 2 clusters (I_1^0 et I_2^0).

Ensuite, l'algorithme détermine 2 nouveaux centres en prenant les centres de gravité des clusters obtenus précédemment (C_1^1 et C_2^1). Ces nouveaux centres induisent une nouvelle partition de l'ensemble des données (I_1^1 et I_2^1) et l'algorithme détermine à nouveau leur centres de gravité (C_1^2 et C_2^2). Ces derniers induisent une nouvelle partition des données (I_1^2 et I_2^2) et le processus se stabilise car l'itération suivante conduit à la même partition. Il est important de remarquer que la partition obtenue finalement dépend du choix initial des centres. En prenant des centres initiaux différents, on pourrait aboutir à une partition des données différente.

Les centres finaux représentent des clusters trouvés et peuvent être utilisés pour prédire les affectations de nouveaux points de données.

Mesure de similarité

Afin de mesurer la similitude entre les points de données et les regrouper correctement, il est courant de calculer la distance entre ces points, en supposant que les points les plus proches les uns des autres sont les plus similaires. La distance euclidienne est souvent utilisée pour ce faire.

Choix du nombre de clusters

Le choix du nombre de clusters k est souvent basé sur des connaissances préalables ou une analyse exploratoire des données. Il peut également être déterminé automatiquement

en utilisant des critères tels que l'indice de silhouette ou la méthode du coude détaillés dans la section [2.2.3](#).

Avantages

La méthode K-means présente plusieurs avantages. Par exemple :

- elle est rapide, robuste et facile à mettre en oeuvre.
- Lorsque les clusters sont distincts ou bien séparés les uns des autres, elle donne de bons résultats.
- Elle est efficace pour traiter les très grandes bases de données car seuls les vecteurs des moyennes sont à conserver en mémoire centrale.
- Sa complexité est linéaire par rapport au nombre d'observations car il n'y a pas de calcul des distances deux à deux des individus.

Cependant, cette méthode présente quelques inconvénients :

Inconvénients

- Cette méthode nécessite de spécifier au préalable le nombre de centres de clusters et puisque les centres initiaux sont choisis de manière aléatoire, il est possible de trouver des solutions différentes à chaque exécution de l'algorithme.
Pour éviter ce problème, il est courant d'utiliser une méthode de choix des centres tels que K-means++. Cette méthode exécute l'algorithme plusieurs fois avec des centres initiaux différents et sélectionne la solution pour laquelle l'inertie intra-cluster est minimale (ou celle pour laquelle l'inertie inter-cluster est maximale).
- Les clusters fournies par cette méthode peuvent dépendre de l'ordre des individus[MacQueen, 1967]
Une solution consiste à mélanger aléatoirement les individus avant de les faire passer pour ne pas être dépendant d'une organisation non maîtrisée des observations.
- Cet algorithme converge souvent vers un minimum(maximum) local de l'inertie intra-cluster (inter-cluster), en fonction du choix des centres initiaux.
- L'influence des valeurs aberrantes sur les résultats est forte
- Cette méthode s'applique uniquement lorsque la moyenne est définie, elle ne s'applique donc pas aux données catégorielles.

Comme mentionné précédemment, K-means ne s'applique qu'aux données continues. Cependant, dans la vie courante, certaines bases de données à analyser ne contiennent que des variables catégorielles et l'algorithme K-means ne peut pas être directement utilisé pour effectuer le partitionnement. Dans ce cas, il peut être intéressant d'utiliser l'algorithme K-modes, qui est une alternative intéressante à K-means pour les données catégorielles. Cette méthode est présentée dans la suite.

2.2.5 K-modes

Le principe de la méthode des K-modes est quasiment identique à celui des K-means mais est adaptée pour les données catégorielles. La différence avec la méthode des K-means

est que la mise à jour des centres des clusters se fait en choisissant le mode de chaque variable catégorielle dans chaque cluster ; tandis que dans la méthode des K-means, elle se fait en calculant la moyenne arithmétique des points de données dans chaque cluster. Cette différence est due à la mesure de distance utilisée pour calculer la similarité entre les points de données.

Principe de base

Les étapes de cet algorithme sont les suivantes :

1. Initialisation : Fixer le nombre de clusters (k) et initialiser les centres de ces clusters aléatoirement.
2. Attribution des points aux clusters : Pour chaque point de données, trouver le cluster le plus proche en utilisant une mesure de dissimilarité adaptée aux variables catégorielles. Assigner le point de données au cluster correspondant.
3. Mise à jour des centres de clusters : Pour chaque cluster, déterminer le mode (la valeur la plus fréquente) de chaque variable catégorielle dans le cluster. Les modes déterminent les nouveaux centres de clusters.
4. Répéter les étapes 2 et 3 jusqu'à ce que la convergence soit atteinte ou qu'un nombre maximum d'itérations soit atteint.

Mesure de similarité

Contrairement à la méthode des K-means, qui utilise la distance euclidienne comme mesure de similarité, la méthode des K-modes utilise une mesure de dissimilarité adaptée aux données catégorielles.

La mesure de dissimilarité la plus couramment utilisée dans la méthode des K-modes est la distance de Hamming. Utiliser cette distance permet ainsi de quantifier la différence entre les valeurs des variables catégorielles pour deux points de données.

La méthode K-modes présente plusieurs avantages et inconvénients [[Huang, 1998](#)].

Avantages

- elle est adaptée aux données catégorielles : la méthode des K-modes est spécifiquement conçue pour les données catégorielles, ce qui la rend plus efficace que la méthode des K-Moyennes pour ce type de données.
- relativement robuste aux valeurs aberrantes et aux données manquantes
- La méthode des K-modes utilise des catégories spécifiques pour définir chaque cluster. Cela signifie que chaque cluster est défini par une combinaison particulière de catégories, plutôt que par des valeurs numériques. Cette caractéristique rend les clusters plus facilement interprétables, car les catégories peuvent avoir un sens intuitif.

Inconvénients

- Sensible à l'initialisation : Comme la méthode des K-Moyennes, la méthode des K-modes est sensible à l'initialisation des centres de clusters et peut produire des résultats différents selon l'initialisation choisie.
- Besoin d'une grande quantité de mémoire : La méthode des K-modes nécessite une grande quantité de mémoire pour stocker la matrice de dissimilarité, surtout lorsque le nombre de catégories est élevé. Cela peut poser des problèmes de mémoire pour des jeux de données de grande taille.

Maintenant que nous avons examiné les méthodes des K-means et des K-modes pour le clustering des données respectivement numériques et catégorielles, explorons à présent une méthode de clustering hybride appelée K-prototypes, qui combine les deux approches pour permettre le clustering de données numériques et catégorielles simultanément.

2.2.6 K-prototypes

Contrairement aux méthodes des K-means et des K-modes, la méthode des K-prototypes utilise à la fois des distances euclidiennes pour les données numériques et des distances de similarité pour les données catégorielles. Cette approche permet d'obtenir des clusters plus précis et mieux adaptés à des jeux de données mixtes.

Principe de base

Les étapes de la méthode K-prototypes sont les suivantes :

1. Initialisation : Comme dans les méthodes K-means et K-modes, K-prototypes commence par fixer un nombre de clusters k et initialiser leurs centres de manière aléatoire.
2. Assignation des observations aux clusters : chaque observation est ensuite assignée au cluster dont le centre est le plus proche, en utilisant une distance mixte qui combine la distance euclidienne pour les variables numériques et une mesure de dissimilarité appropriée pour les variables catégorielles.
3. Mise à jour des centres des clusters : une fois que toutes les observations ont été assignées à un cluster, les centres de chaque cluster sont mis à jour en calculant la moyenne des observations qui lui ont été assignées pour les variables numériques et en déterminant les modes (valeurs les plus fréquentes) pour les variables catégorielles.
4. Les étapes 2 et 3 sont répétées jusqu'à ce que les centres des clusters ne changent plus ou jusqu'à ce qu'un nombre maximum d'itérations soit atteint.

Mesure de similarité entre individus

La mesure de similarité dans le cas de la méthode K-prototype est détaillée dans [Huang, 1998]. Nous nous sommes inspirés de cet article pour rédiger cette partie.

La similarité entre deux individus dans la méthode K-prototypes, prend en compte à la fois la distance euclidienne pour les variables numériques et la distance de Hamming pour les variables catégorielles. Une fois calculées, il est nécessaire de combiner ces distances

pour obtenir une mesure de distance globale entre les deux individus. Pour cela, on peut utiliser une pondération qui permet de donner plus ou moins d'importance aux variables catégorielles par rapport aux variables numériques.

Par exemple, pour deux individus X et Y , la distance couramment utilisée dans la méthode K-prototype est donnée par la formule suivante où le premier terme sous le radical représente la distance euclidienne utilisée pour les variables numériques et le second terme représente la distance de Hamming utilisée pour les variables catégorielles, pondéré par un facteur :

$$d_{K-prototypes}(X, Y) = \sqrt{\sum_{j=1}^p (x_j - y_j)^2 + \gamma \sum_{j=p+1}^m \delta(x_j, y_j)}$$

où :

- m est le nombre de variables totales (numériques et catégorielles)
- p est le nombre de variables numériques
- $\delta(x_i, y_i)$ est la fonction delta de Kronecker qui prend la valeur 1 si $x_i = y_i$ et 0 sinon.
- γ est un coefficient de pondération pour les attributs catégoriels.

On peut remarquer que cette distance prend en compte à la fois la distance euclidienne entre les attributs numériques et la distance de Hamming entre les attributs catégoriels. Le poids γ (qui varie entre 0 et 1) est utilisé ici pour éviter de favoriser l'un ou l'autre type d'attribut (numérique ou catégoriel) dans le calcul de cette distance totale entre deux objets mixtes. En d'autres termes, il permet de pondérer l'importance relative des deux termes dans le calcul de la distance totale.

Une petite valeur de γ indique que le regroupement est dominé par les attributs numériques, tandis qu'une grande implique que les attributs catégoriels dominent le regroupement.

En introduisant le poids γ , on s'assure que les attributs catégoriels ne sont pas négligés dans le processus de clustering, et que les deux types d'attributs ont une influence équilibrée sur la distance totale.

Il est important de noter que le choix de ce poids peut affecter les résultats du clustering. Par conséquent, il est souvent utile de tester différentes valeurs de poids pour déterminer l'impact de chaque choix sur les résultats de l'analyse. Par exemple, dans son étude, Huang [Huang, 1998] utilise une approche où il fixe différentes valeurs de gamma pour évaluer les résultats de l'algorithme K-prototypes sur un jeu de données. Tout d'abord, il normalise les variables numériques et calcule la moyenne de leurs écart-types. Ensuite, il expérimente différentes valeurs de gamma autour de cette moyenne. Pour évaluer les performances du modèle obtenu pour chaque valeur de gamma, il utilise une mesure de précision (l'accuracy). Le meilleur modèle est celui qui obtient le score d'accuracy le plus élevé.

Avantages

La méthode K-prototypes présente plusieurs avantages. Par exemple :

L'adaptabilité aux données mixtes : La méthode K-prototypes est conçue pour traiter les données mixtes, qui combinent des variables numériques et catégorielles. Cela permet de traiter des ensembles de données plus variés et plus complexes, qui ne peuvent pas être traités par des méthodes de clustering traditionnelles.

Inconvénients

- La sélection aléatoire des centres de clusters initiaux entraîne l'incertitude et le caractère aléatoire des résultats de clustering, et le nombre de clusters k doit être déterminé manuellement.
- la distance de Hamming est utilisée pour calculer la dissimilarité entre les données catégorielles et les centres de clusters, ce qui entraîne une perte d'informations et l'incapacité de refléter objectivement la situation réelle entre les objets de données et les clusters, ce qui entraîne parfois des résultats de clustering inexacts.
- le paramètre utilisé pour ajuster la proportion entre les données catégorielles et les données numériques doit être déterminé manuellement.

2.2.7 Classification ascendante hiérarchique

Les méthodes de clustering basées sur les centroïdes évoquées précédemment sont extrêmement populaires mais peuvent être mal appliquées aux ensembles de données du monde réel. Car en plus d'être sensibles aux valeurs aberrantes, il est souvent difficile de déterminer le nombre approprié de centroïdes k à spécifier à l'avance. Dans la pratique, pour contourner ces lacunes, il est courant d'explorer un autre type d'algorithme de clustering : le clustering hiérarchique. Un exemple de clustering hiérarchique est la classification ascendante hiérarchique (CAH) [Lebart *et al.*, 1995].

La Classification Ascendante Hiérarchique (CAH) vise à regrouper des données selon leur ressemblance, en construisant un arbre de classification appelé dendrogramme. Elle est dite "ascendante" car elle part des individus pour construire progressivement des groupes de plus en plus larges et "hiérarchique" car elle crée des sous-groupes inclus dans des groupes plus vastes.

Principe de base

Le principe consiste à regrouper des individus selon un critère de ressemblance préalablement défini, indiquant la proximité ou l'éloignement entre chaque paire d'individus. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante. Les étapes sont les suivantes :

- Tout d'abord, chaque observation est considérée comme un cluster individuel.
- Ensuite, les paires de clusters les plus similaires sont fusionnées pour former un nouveau cluster.
- Fusion des deux clusters les plus proches en un nouveau cluster.
- Les étapes 2 et 3 sont répétées jusqu'à ce qu'un seul cluster contenant toutes les observations soit créé.

La figure suivante illustre ces étapes. Elle est extraite de [Lebart *et al.*, 1995]

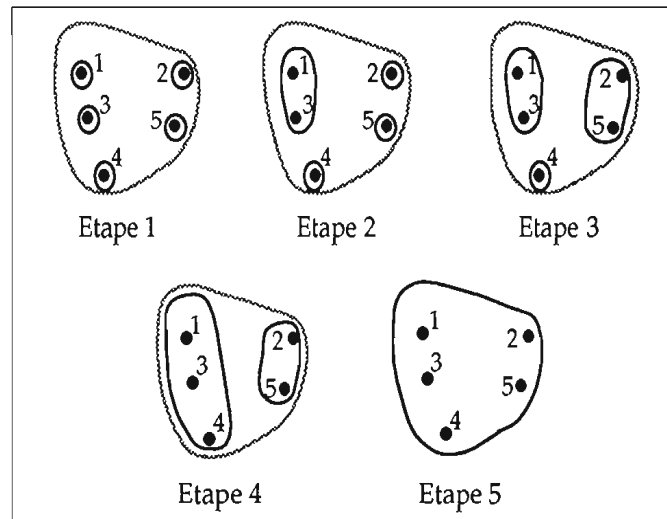


FIGURE 2.9 – Classification ascendante hiérarchique appliquée sur 5 points

Les résultats peuvent être visualisés sous forme de dendrogramme, qui est une représentation graphique de la manière dont les clusters ont été fusionnés. Un seuil de distance est choisi pour couper le dendrogramme et obtenir un nombre souhaité de clusters.

Mesure de similarité entre individus

De la même manière que pour le clustering basé sur le centroïde, il est important de définir explicitement quand est ce que deux individus sont proches. Lorsque chaque individu est son propre cluster, la distance euclidienne est la mesure la plus couramment utilisée pour calculer la similarité entre deux individus. Elle a été définie dans la section 2.2.1. Les individus les plus proches sont ceux pour lesquels cette distance est minimale.

Stratégies d'agrégation : mesures de dissimilarité entre les clusters

Il existe plusieurs mesures de dissimilarité entre clusters. Les plus couramment utilisées sont le saut minimum, le saut maximum et la distance de Ward.

1. Le saut minimum ou «single linkage» mesure la distance entre un individu et un groupe en déterminant la distance minimale entre l'individu et tous les membres du cluster. Si on désigne par $C1$ et $C2$ les deux clusters représentés sur la figure 2.10,

$$d(C1, C2) = \min_{a \in C1, b \in C2} d(a, b)$$

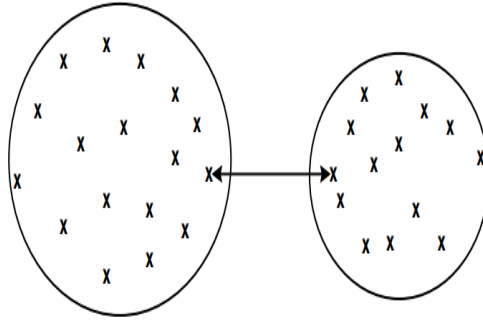


FIGURE 2.10 – Illustration du saut minimum

2. Le saut maximum ou «complete linkage» mesure la distance entre un individu et un groupe en déterminant la distance maximale entre l'individu et tous les membres du cluster. Si on désigne par $C1$ et $C2$ les deux clusters représentés sur la figure 2.11,

$$d(C1, C2) = \max_{a \in C1, b \in C2} d(a, b)$$

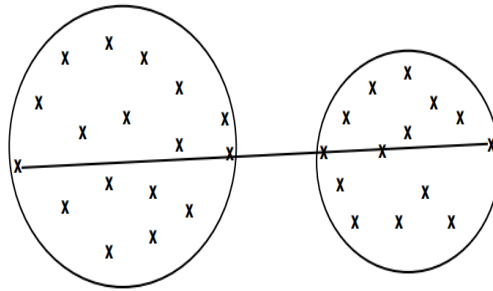


FIGURE 2.11 – Illustration du saut maximum

3. La distance de Ward ou «Ward linkage» mesure la distance entre un individu et un groupe en utilisant une analyse de la variance pour déterminer la distance minimale entre l'individu et tous les membres du cluster. En appliquant la distance de Ward sur le graphique précédent, on aurait :

$$d^2(C_1, C_2) = \frac{n_1 \times n_2}{n_1 + n_2} d^2(G_1, G_2)$$

où G_1 et G_2 sont les centres de gravité des clusters C_1 et C_2 et n_1, n_2 , leur nombre d'observations respectifs.

En général, le choix de la mesure de dissimilarité dépend de la nature des données et des objectifs de la classification hiérarchique. La distance de Ward est néanmoins la plus utilisée dans la pratique car ce critère maximise l'inertie inter-cluster lorsqu'on utilise la distance euclidienne

Avantages

- Pas besoin de spécifier le nombre de clusters : La CAH n'exige pas la spécification préalable du nombre de clusters. Elle produit un dendrogramme qui est un diagramme arborescent où les clusters sont représentés par des nœuds et les liens entre les nœuds représentent les similarités entre les clusters. Il offre ainsi une représentation visuelle des regroupements possibles des données. Par exemple, si nous souhaitons obtenir 3 clusters, nous pouvons couper le dendrogramme de manière à séparer les données en trois groupes distincts (pour une visualisation du dendrogramme, voir le chapitre 3).
- Interprétabilité : La CAH produit une structure de clusters hiérarchique qui peut être visualisée sous forme de dendrogramme, ce qui facilite l'interprétation des résultats.
- La CAH est relativement robuste aux valeurs aberrantes et aux erreurs de mesure.

Inconvénients

- La qualité des résultats de la CAH dépend fortement des mesures de distance ou de similarité choisies. Des mesures inappropriées peuvent entraîner des regroupements inefficaces ou erronés.
- La CAH peut être coûteuse en temps de calcul, surtout pour un grand nombre d'éléments de données car elle nécessite de calculer la distance entre chaque paire d'éléments de données, ce qui peut être très coûteux en termes de temps de calcul pour les ensembles de données volumineux.
- Les dendrogrammes peuvent être difficiles à lire pour les grandes structures de clusters hiérarchiques, ce qui peut compliquer l'interprétation des résultats.

Les techniques présentées ci-dessus ont été mises en oeuvre sur la base de donnée contenant les patients ayant demandé un traitement pour la première fois au cours de l'année 2021. Les résultats obtenus sont présentés dans le chapitre suivant.

Chapitre 3

Résultats

3.1 Analyse en clusters des consommateurs de cannabis ayant demandé un traitement pour la première fois au cours de l'année 2021 en Belgique

Dans cette section, nous nous intéressons aux patients remplissant critères suivants :

- Critère n°1 : Le cannabis constitue la substance principale pour laquelle ils ont demandé un traitement.
- Critère n°2 : La demande de traitement a été faite au cours de l'année 2021.
- Critère n°3 : Ils n'ont jamais suivi de traitement auparavant.

Nous avons identifié **1847** patients qui satisfont à ces critères. Commençons par présenter les statistiques descriptives des variables du jeu de données.

3.1.1 Analyse exploratoire des données

Analyse des caractéristiques socio-démographiques des patients

L'analyse du jeu de données fait ressortir le fait que les patients sont relativement jeunes. L'âge moyen est de 25 ans (min= 13, max= 64), avec un écart-type de 9.04. L'âge moyen lors du premier usage du cannabis est de 15 ans (min= 7, max=51), écart-type de 3.44. la figure 3.1 permet de visualiser la distribution des âges.



FIGURE 3.1 – Distribution de l'âge des patients et leur âge lors du premier usage du cannabis

Un rapport détaillé de la répartition des patients en fonction des caractéristiques socio-démographiques est contenu dans le tableau [A.1](#) en annexe. On peut y remarquer que parmi les patients, 19.92% sont des femmes tandis que 80.08% sont des hommes. Un peu moins de la moitié des patients (46.51%) ont déclaré avoir vécu avec un parent au cours des 30 derniers jours. Ensuite, 17.27% des patients sont en couple, 25.99% vivent seuls et des proportions plus faibles vivent avec des amis (4.22%) ou avec leur famille (5.14%).

La grande majorité des patients (89.55%) ont un domicile fixe. Un nombre plus faible de patients résident dans une institution (3.95%), dans des logements variables (3.74%) ou en situation de rue (0.87%). La proportion des patients en prison est également faible (1.35%).

En ce qui concerne le niveau d'instruction, nous observons que la majorité des patients ont un niveau d'instruction secondaire (63.18%) ou primaire (28.05%). Seulement une petite proportion a un niveau d'instruction supérieur (6.01%). 1 patient sur 3 déclare ne pas avoir de source de revenus, tandis que 29,99% ont un revenu provenant d'un salaire. D'autres sources de revenus comprennent les allocations familiales (2,38%), les indemnités (12.89%), le CPAS (7.58%). La majorité des patients sont soit étudiants/écoliers (36.33%) ou au chômage (28.48%).

Analyse du profil d'addiction des patients

De la figure 3.2, on observe que les principales sources de renvoi en traitement sont la justice (40.8%), les renvois individuels (24.42%) ou la famille (9.58%). Des proportions plus faibles de patients sont référés par des amis (2.33%), des centres pour toxicomanes (1.03%), des hôpitaux (4.33%) ou des médecins (4.71%).

Près de deux tiers des patients (59.1%) consomment le cannabis quotidiennement. Une proportion significative utilise des substances de manière moins fréquente, avec 10.40% déclarant une utilisation de 2 à 3 jours par semaine, 7.74% de 4 à 6 jours par semaine et 7.26% moins d'une fois par semaine.

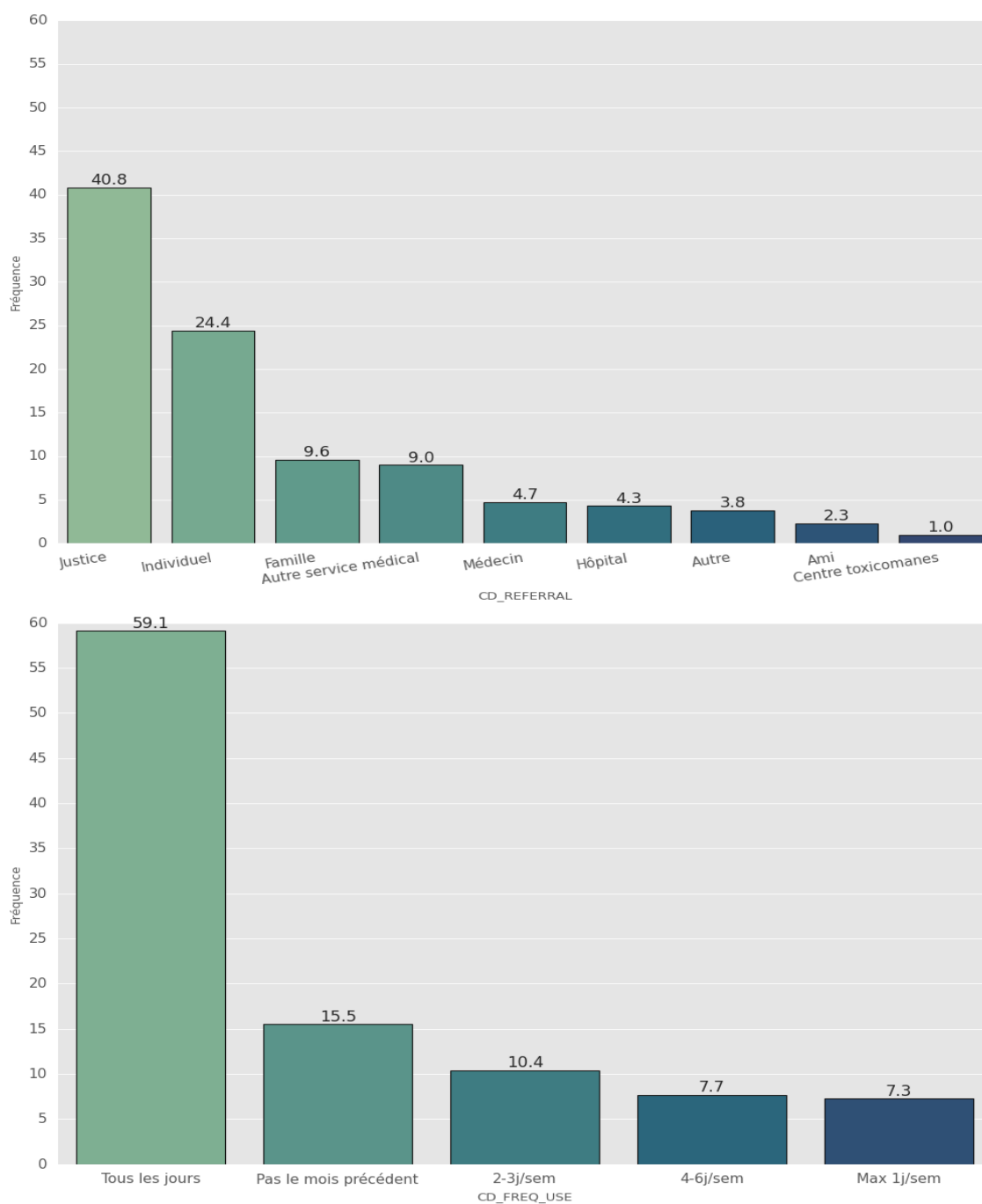


FIGURE 3.2 – Répartition de l'origine du renvoi en traitement et de la fréquence de consommation du cannabis chez les patients

Le tableau A.2 en annexe A présente la répartition des patients en fonction du profil d'addiction. On peut y remarquer que la grande majorité des patients (98.97%) n'ont jamais eu recours aux substances psychoactives par injection.

Type de substances utilisées par les patients

Le tableau 3.1 met en évidence les types de substances couramment utilisées, ainsi que les moins répandues. Selon les types de substances, on peut faire les observations suivantes :

- Opiacés : La majorité des patients (99.03%) n'utilisent pas d'opiacés. Une petite proportion (0.65%) déclare consommer l'héroïne. Quant à la méthadone, aucun des patients n'en consomme. Les autres opiacés sont peu fréquents, avec seulement 0.11% des patients déclarant en consommer.
- Cocaïne : La cocaïne est consommée par 9.64% des patients. Parmi ceux-ci, 5.63% consomment de la cocaïne en poudre, tandis qu'une plus petite proportion consomme du crack (0.54%) ou d'autres formes de cocaïne (0.05%).
- Stimulants : En dehors de la cocaïne, d'autres stimulants sont également utilisés par les patients; environ 6.82% des patients consomment des stimulants autres que la cocaïne. Parmi ces stimulants, les amphétamines sont les plus courantes (3.95%), suivies de la MDMA (3.14%) et de la mephédrone (0.70%).
- Hypnotiques : Une petite proportion des patients (2.38%) déclare consommer des hypnotiques, tels que les barbiturates (1.73%) ou les benzodiazépines (1.73%).
- Hallucinogènes : Les hallucinogènes sont consommés par 3.03% des patients. Parmi ceux ci, la ketamine est la plus consommée (2.06%), suivie du LSD (0.92%) et d'autres hallucinogènes (0.43%).
- Inhalants volatils : La consommation d'inhalants volatils est rare parmi les patients, avec seulement 0.22% déclarant en consommer.
- Cannabis : La marijuana est la forme de cannabis la plus répandue (57.50%), suivie du haschisch (5.52%).
- Alcool : Environ 12.67% des patients déclarent consommer de l'alcool, tandis que la majorité (87.33%) n'en consomme pas.

TABLE 3.1 – Répartition des patients en fonction du type de substances utilisées

Types de substances	variables	effectif	fréquence(%)
Opiacés	FL_OPIATES	Non	1829
		Oui	18
	FL_HEROIN	Non	1835
		Oui	12
	FL_METHADONE	Non	1847
	FL_BUPRENORPHINE	Non	1847
	FL_FENTANYL	Non	1847
	FL_OTHER_OPIATES	Non	1845
		Oui	2
Cocaïne			

Stimulants (autres que la cocaïne)	FL_COCAINE	Non	1669	90.36
		Oui	178	9.64
	FL_POWDER_COCAINE	Non	1743	94.37
		Oui	104	5.63
	FL_CRACK	Non	1837	99.46
		Oui	10	0.54
	FL_OTHER_COCAINE	Non	1846	99.95
		Oui	1	0.05
	FL_STIMULANTS	Non	1721	93.18
		Oui	126	6.82
	FL_AMPHETAMINES	Non	1774	96.05
	FL_AMPHETAMINES	Oui	73	3.95
	FL_METHAMPHETAMINES	Non	1844	99.84
		Oui	3	0.16
Hypnotiques	FL_MDMA	Non	1789	96.86
		Oui	58	3.14
	FL_MEPHEDRONE	Non	1834	99.30
		Oui	13	0.70
	FL_OTHER_STIMULANTS	Non	1846	99.95
		Oui	1	0.05
	FL_HYPNOTICS	Non	1803	97.62
		Oui	44	2.38
	FL_BARBITURATES	Non	1847	100.00
	FL_BENZODIAZEPINES	Non	1815	98.27
Hallucinogènes		Oui	32	1.73
	FL_GHB	Non	1839	99.57
	FL_GHB	Oui	8	0.43
	FL_OTHER_HYPNOTICS	Non	1847	100.00
	FL_HALLUCINOGENS	Non	1791	96.97
		Oui	56	3.03
	FL_LSD	Non	1830	99.08
	FL_LSD	Oui	17	0.92
	FL_KETAMINE	Non	1809	97.94
		Oui	38	2.06
Inhalants volatils	FL_OTHER_HALLUCINOGENS	Non	1839	99.57
		Oui	8	0.43
	FL_VOLATILE	Non	1843	99.78
		Oui	4	0.22

Cannabis	FL_MARIJUANA	Non	785	42.50
		Oui	1062	57.50
	FL_HASH	Non	1745	94.48
		Oui	102	5.52
	FL_OTHER_CANNABIS	Non	1839	99.57
		Oui	8	0.43
Alcool	FL_ALCOHOL	Non	1613	87.33
		Oui	234	12.67
Autres types de substances	FL_OTHER_SUBSTANCE	Non	1847	100.00
		Oui		

Comme le montre la figure 3.3, près de 9 patients sur 10 (91.7%) ont au plus deux substances problématiques et une faible proportion (0,9%) a entre 6 et 10 substances problématiques.

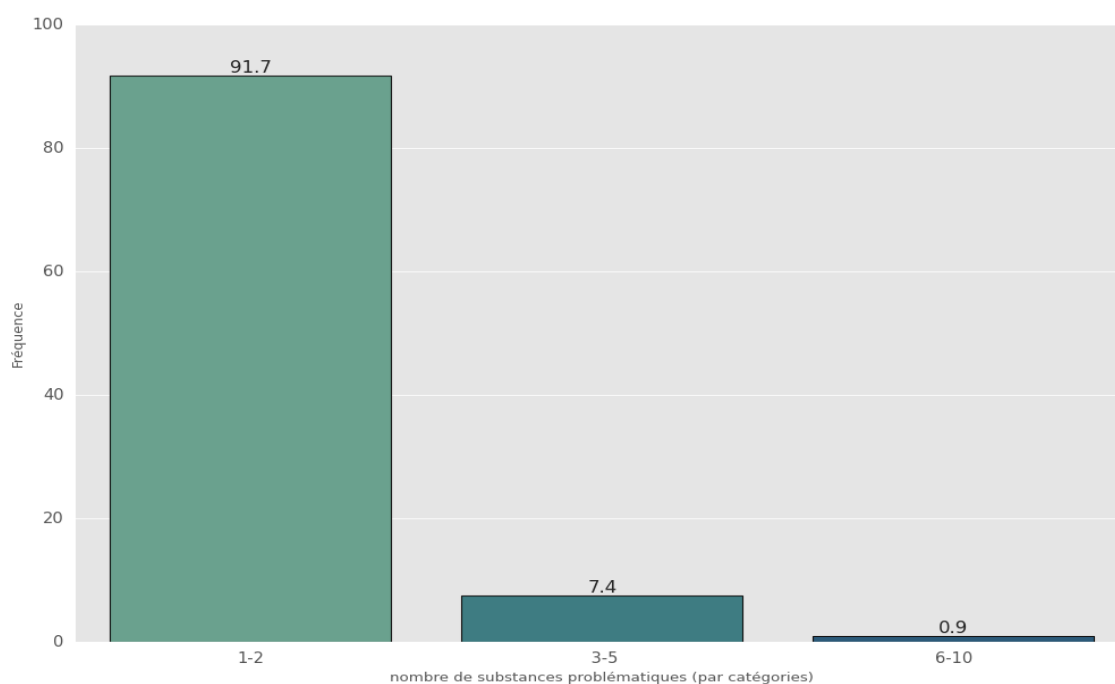


FIGURE 3.3 – Répartition du nombre total de substances problématiques par catégories

Après avoir présenté les statistiques descriptives qui nous donnent un aperçu de la distribution et des caractéristiques générales du jeu de données, il est maintenant temps d'explorer davantage la structure sous-jacente en utilisant des méthodes de clustering.

3.1.2 Clustering

Dans le but d'identifier les clusters parmi les patients, nous explorons dans cette section différentes méthodes de clustering afin de déterminer celle qui fournit la meilleure partition. La meilleure partition est identifiée à l'aide du score de silhouette moyen (voir la section 2.2.3 pour des détails sur cette métrique). Plus ce score est élevé, plus la partition est meilleure.

Dans le but de traiter les données d'entrée (qui sont mixtes) et de capturer la variance totale qui y est contenue, nous avons appliqué une AFDM. Les 69 premiers axes factoriels permettaient de capter l'information contenue dans les données et celles-ci ont été utilisées comme variables d'entrée pour l'application des méthodes K-means et CAH.

K-means

La mesure de similarité utilisée pour appliquer cette méthode est la distance euclidienne. La détermination du nombre optimal de clusters a été effectuée en faisant varier le nombre de centres initiaux k entre 2 et 7. Les scores de silhouette moyens obtenus sont contenus dans le tableau 3.2.

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
score de silhouette moyen	0.372	0.048	0.022	0.031	0.033	0.073

TABLE 3.2 – Kmeans : score de silhouette moyen en fonction du nombre de clusters

En examinant les scores de silhouette moyens dans le tableau 3.2, on observe que le score de silhouette moyen le plus élevé vaut 0.372 et est obtenu pour $k=2$.

Pour $k=3$, ce score est très faible (seulement 0.048) ce qui suggère une séparation insuffisante des clusters. Le même constat est fait pour $k=4$. Le score de silhouette augmente légèrement pour $k=5$ à 0.031, mais reste assez bas. La séparation des clusters reste donc également insuffisante dans ce cas. Pour $k=7$, le score augmente légèrement par rapport aux valeurs précédentes (0.073).

Il apparaît donc qu'en considérant la méthode K-means, le meilleur nombre de clusters semble être 2, car il obtient le score de silhouette moyen le plus élevé. Les autres nombres présentent des scores de silhouette relativement faibles, indiquant une séparation médiocre entre les clusters¹.

La figure 3.4 permet de visualiser le tracé de silhouette et la partition obtenue.

1. Nous avons testé des valeurs de k au delà de 7 clusters (8 à 15) et avons observé de plus petites valeurs du score moyen de silhouette.

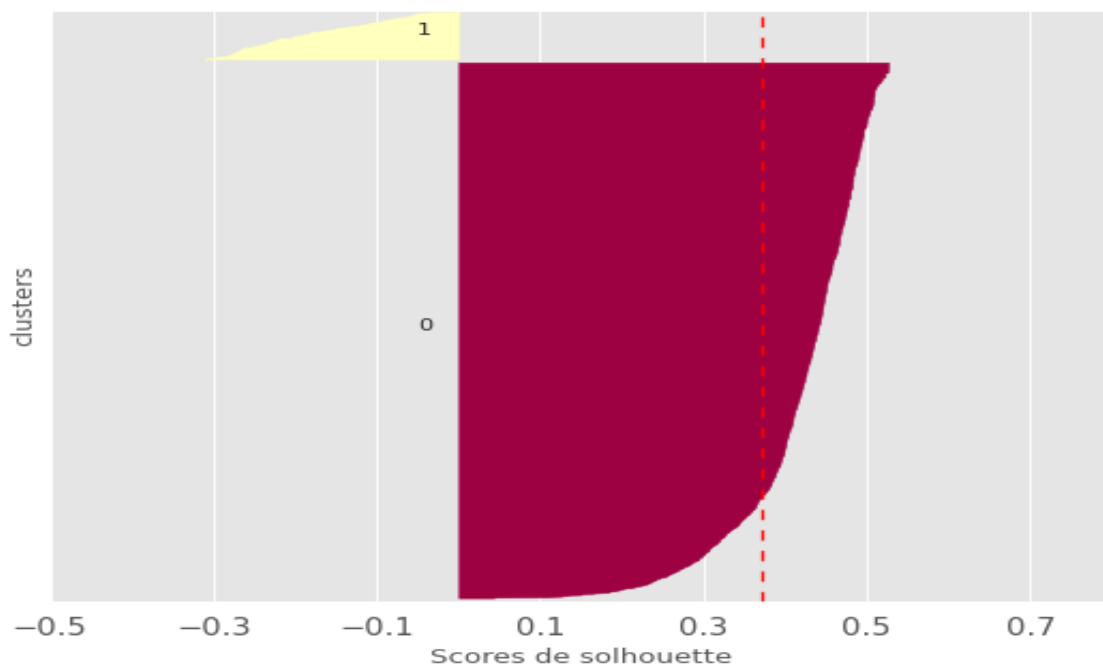


FIGURE 3.4 – Analyse des tracés de silhouette (pour $k = 2$)

La deuxième méthode utilisée en vue de comparer les différents clusters est la classification ascendante hiérarchique.

Classification ascendante hiérarchique

La classification ascendante hiérarchique a également été implémentée sur les axes factoriels issus de l'AFDM, en prenant la distance euclidienne comme mesure de similarité. Le tableau 3.3 fournit les scores de silhouette moyens obtenus pour un nombre de clusters allant de 2 à 7.

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
score de silhouette moyen	0.448	0.447	0.438	0.417	0.028	0.031

TABLE 3.3 – CAH : score de silhouette moyen en fonction du nombre de clusters

En examinant ce tableau, les scores de silhouette moyen sont élevés pour $k=2$ et $k=3$ (0.448 et 0.447 respectivement), indiquant une meilleure séparation des clusters et une meilleure cohésion à l'intérieur de chaque cluster. Pour $k=4$ et $k=5$, les scores de silhouette moyen diminuent légèrement (0.438 et 0.417 respectivement), indiquant une séparation moins marquée par rapport aux valeurs précédentes. Pour $k=6$ et $k=7$, les scores de silhouette moyens sont très faibles (0.028 et 0.031 respectivement), suggérant une mauvaise séparation des clusters et une cohésion très faible.

Il apparaît donc qu'en considérant la méthode CAH, le meilleur nombre de clusters semble être 2, car il obtient le score de silhouette moyen le plus élevé. La figure 3.5 permet de visualiser le tracé de silhouette et la partition obtenue.

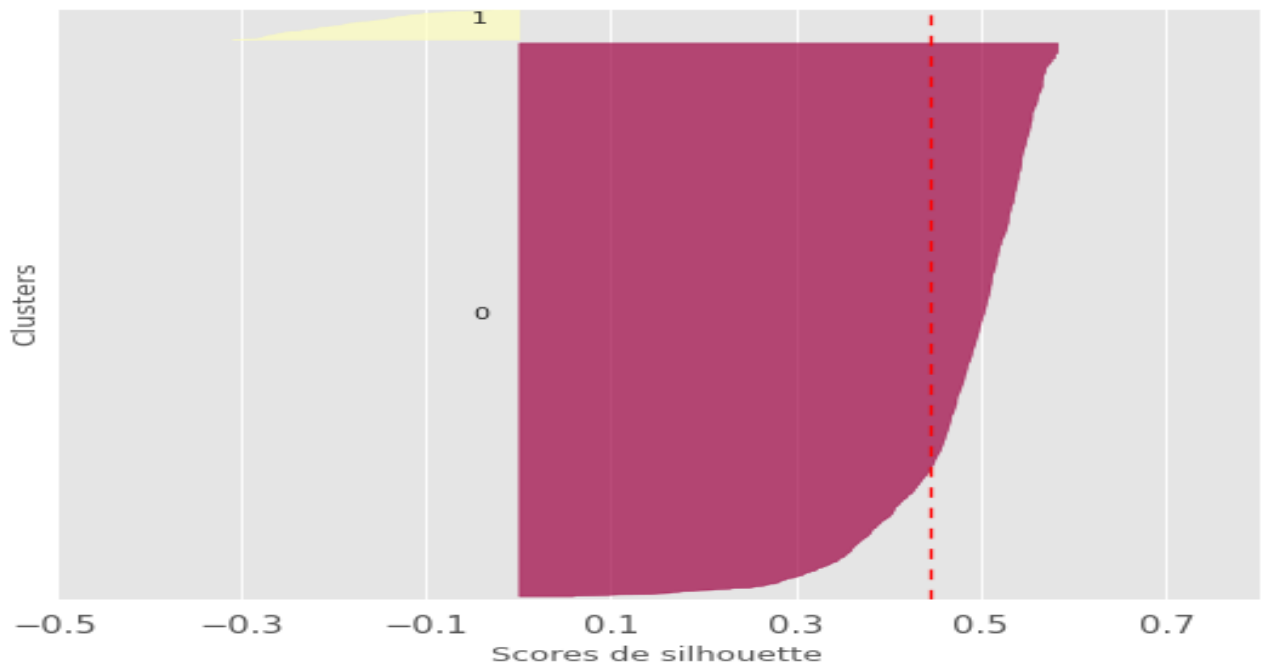


FIGURE 3.5 – Analyse des tracés de silhouette (pour $k = 2$)

Dans la suite, nous présentons la partition fournie par la méthode K-modes

K-modes

Puisque la méthode K-modes fonctionne avec des données catégorielles uniquement, il était nécessaire de discrétiser les âges avant de l'appliquer. Nous avons défini chaque intervalle de manière à inclure la borne supérieure tout en excluant la borne inférieure. Deux nouvelles variables qui correspondent aux classes d'âges ont donc été créées. La figure 3.6 permet de visualiser les proportions au sein de chaque classe.

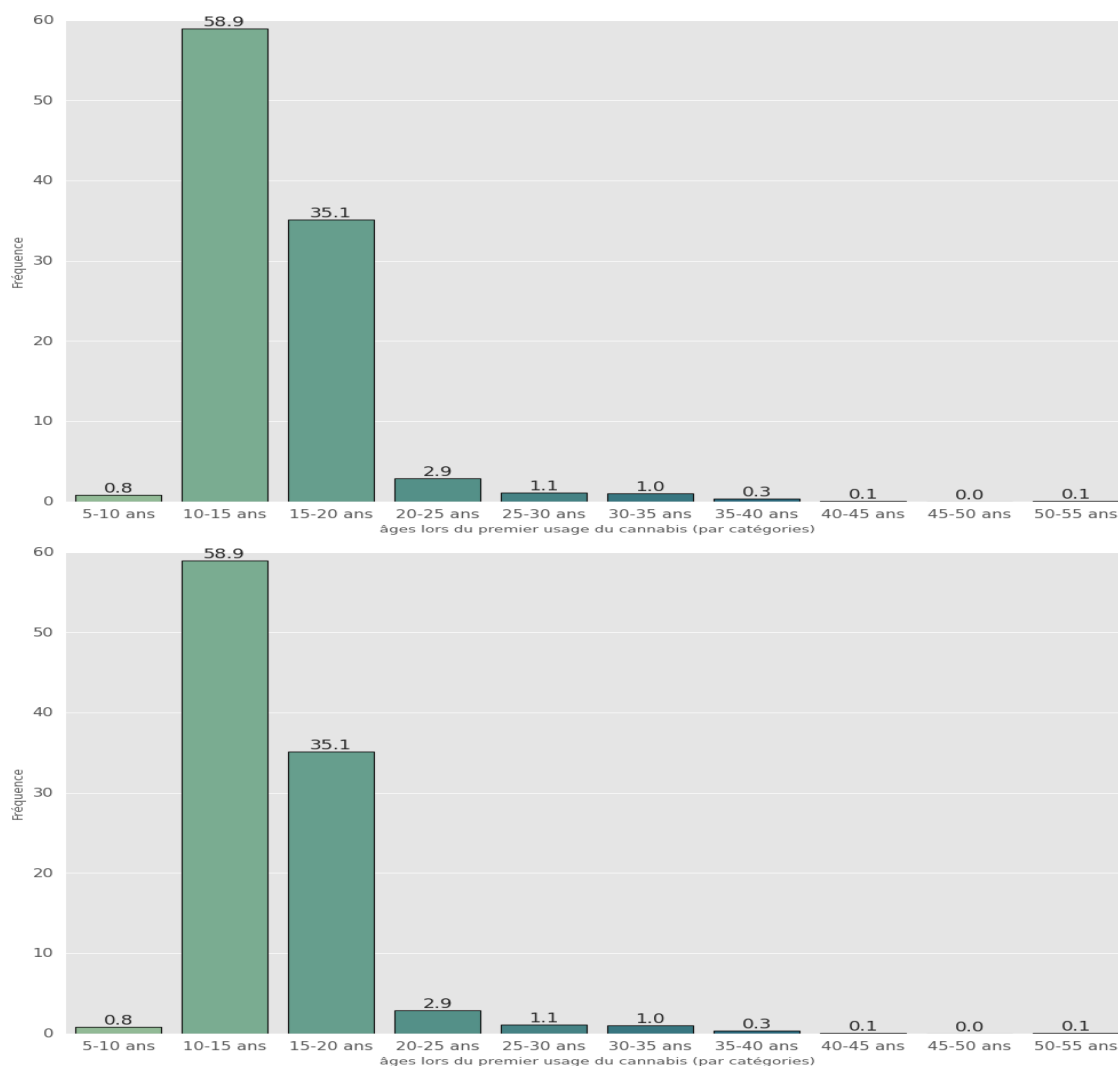


FIGURE 3.6 – Répartition des âges (catégories)

Sur la figure 3.6, on observe que 1 patient sur 3 (30.5%) est âgé entre 15 et 20 ans. Les patients âgés de 60 ans et plus sont les moins nombreux (0.1%). Près de 2 patients sur 3 (58.59%) ont fait usage du cannabis pour la première fois entre 10 et 15 ans. Ces observations sont en droite ligne avec les analyses faites précédemment et confirment que les patients demandant un traitement pour une dépendance au cannabis en Belgique sont relativement jeunes.

Ces nouvelles variables ont été combinées aux autres variables catégorielles afin d'appliquer la méthode K-modes. Le tableau 3.4 ci-dessous présente les scores moyens obtenus.

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
score de silhouette moyen	0.040	0.073	-0.006	-0.033	-0.076	-0.043

TABLE 3.4 – score de silhouette moyen en fonction du nombre de clusters : Kmodes

Le tableau 3.4 montre que le score de silhouette moyen augmente légèrement de 0.040

pour $k=2$ à 0.073 pour $k=3$, indiquant une amélioration de la qualité des clusters. Cependant, à partir de $k=4$, le score de silhouette moyen devient négatif, ce qui suggère un chevauchement et une séparation moins claire entre les clusters. Cette tendance se maintient pour $k=5$, $k=6$ et $k=7$.

Ainsi, selon la méthode *k*modes, le meilleur résultat semble être obtenu pour $k=3$, où le score de silhouette moyen est le plus élevé. La figure 3.7 permet de visualiser le tracé de silhouette ainsi que la partition obtenue.

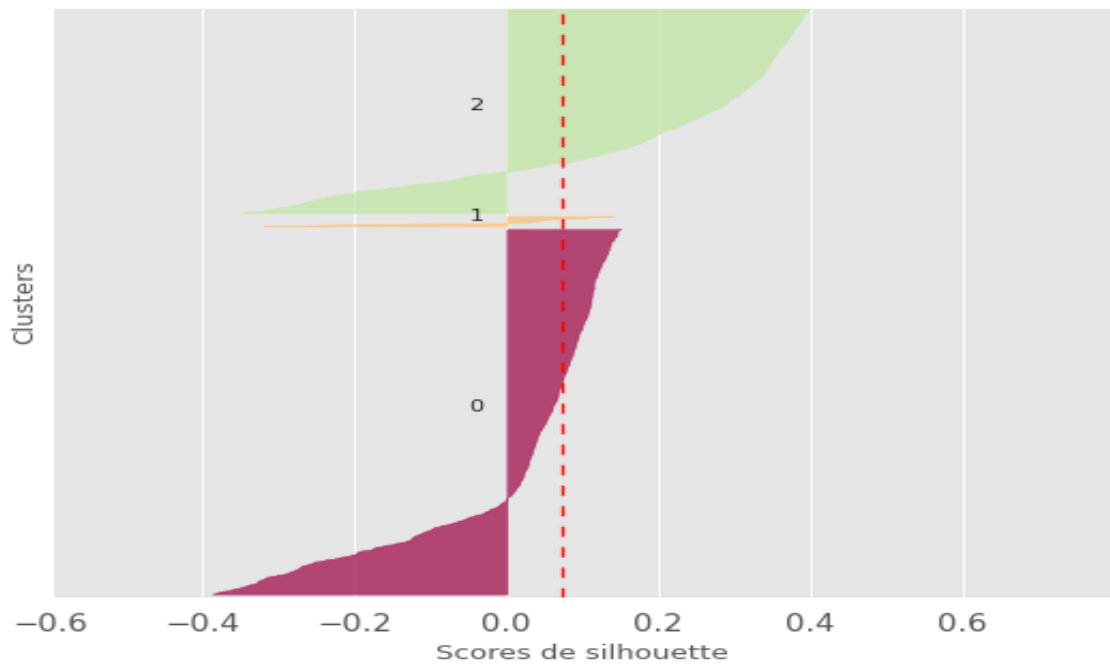


FIGURE 3.7 – Analyse des tracés de silhouette (pour $k = 3$)

Dans la suite, nous présentons le résultat fourni par la méthode K-prototypes.

K-prototypes

Le tableau 3.5 ci-dessous présente le score de silhouette moyen en fonction du nombre de clusters. On y observe que le score de silhouette maximal est 0.324, obtenu pour $k=2$. Cela suggère que, sur la base du score moyen de silhouette, une partition en 2 clusters est optimale.

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
score de silhouette moyen	0.324	0.299	0.215	0.148	0.154	0.138

TABLE 3.5 – score de silhouette moyen en fonction du nombre de clusters : Kprototypes

En examinant les résultats du tableau 3.5, on observe que le score de silhouette moyen le plus élevé est obtenu lorsque 2 clusters sont utilisés, ce qui suggère une séparation raisonnable des clusters. Lorsque trois clusters sont utilisés ($k=3$), le score de silhouette

moyen diminue légèrement à 0.299. Lorsque le nombre de clusters augmente à 4 ($k=4$), le score de silhouette moyen diminue davantage à 0.215. Cette diminution suggère une moins bonne séparation des instances entre les clusters et une cohésion plus faible à l'intérieur des clusters. Lorsque cinq ($k=5$), six ($k=6$) et sept ($k=7$) clusters sont utilisés, les scores de silhouette moyens continuent de diminuer, atteignant respectivement 0.148, 0.154 et 0.138. Ces résultats indiquent une séparation moins satisfaisante des patients entre les clusters et une cohésion relativement faible à l'intérieur des clusters. Ainsi, en utilisant la méthode K-prototypes, les résultats montrent une tendance à une meilleure qualité de regroupement avec 2 clusters. La figure 3.8 permet de visualiser le tracé de silhouette ainsi que la partition obtenue.

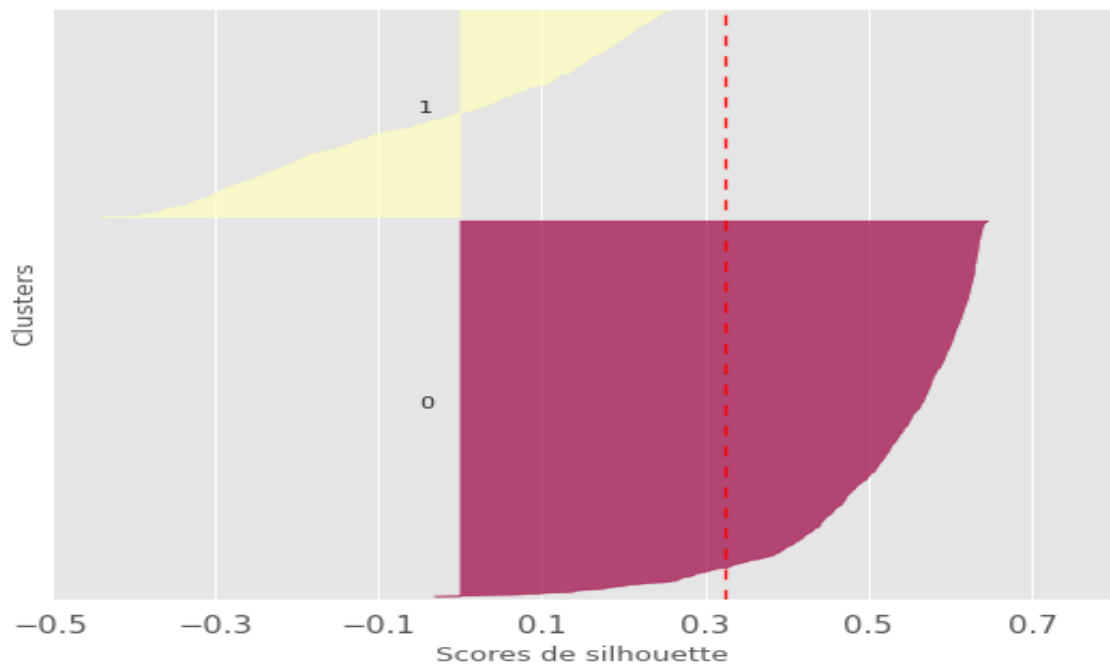


FIGURE 3.8 – Analyse des tracés de silhouette (pour $k = 2$)

3.1.3 Comparaison des scores de silhouette issus des méthodes K-means, CAH, K-modes, K-prototypes et sélection de la meilleure méthode pour notre jeu de donnée

Nous avons précédemment identifié la meilleure partition fournie par chaque méthode. A présent, nous comparons les différents scores de silhouette moyen obtenus afin de retenir la meilleure d'entre toutes pour la suite des analyses. Pour cela, le graphique 3.9 permet de visualiser et de comparer les performances des différentes méthodes sur la base du score de silhouette moyen obtenu.

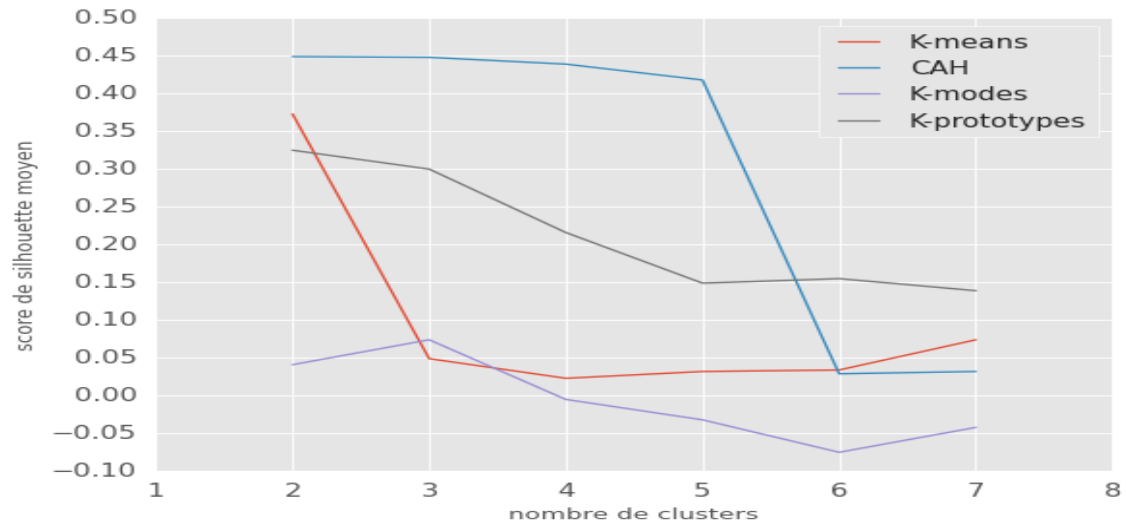


FIGURE 3.9 – Scores de silhouette moyen en fonction du nombre de clusters

Nous avons finalement choisi la CAH car en examinant le graphe 3.9, il apparaît que, la CAH présente globalement les valeurs les plus élevées du score de silhouette moyen pour la plupart des nombres de clusters fixés. Cela signifie que les clusters obtenus par cette méthode sont mieux séparés et compacts que ceux obtenus à partir des méthodes K-means, K-modes et K-prototypes. On note également que le nombre optimal de clusters dans les données est deux(2) car ce nombre obtient le score le plus élevé.

L'analyse en cluster des patients ayant demandé un traitement pour une dépendance au cannabis et dont la demande a été faite pour la première fois au cours de l'année 2021 a donc permis d'identifier 2 clusters. Comme on peut le voir sur la figure 3.10, le dendrogramme de la CAH met en évidence deux branches bien distinctes, illustrant l'existence de deux profils distincts de patients.

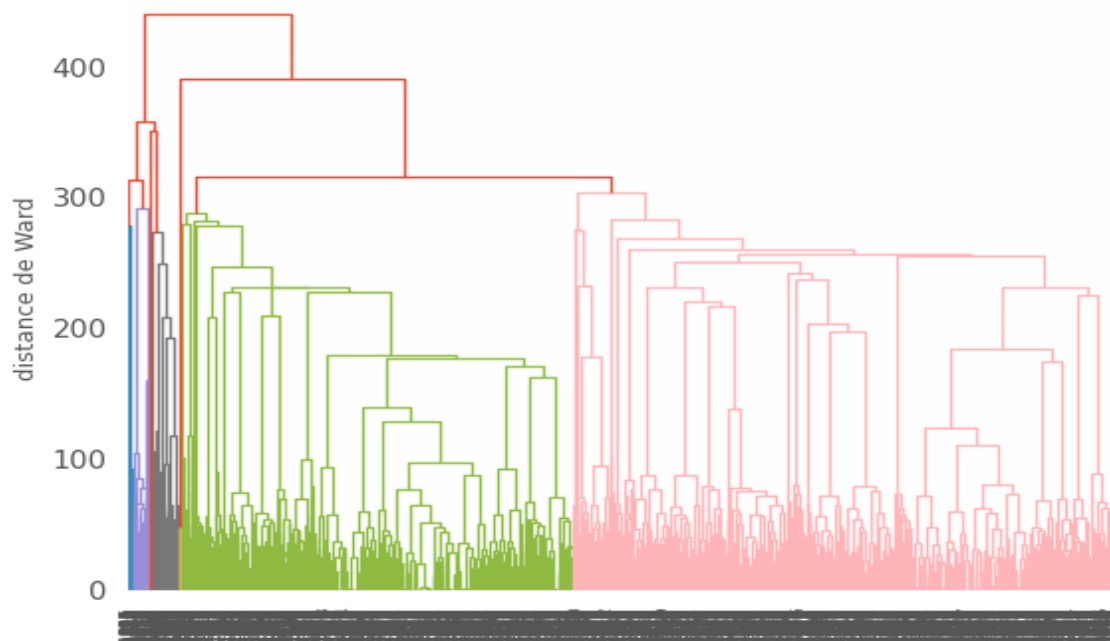


FIGURE 3.10 – Dendrogramme de la classification ascendante hiérarchique

Les proportions obtenues dans les clusters suggèrent que la majorité des patients présentent un profil similaire. En effet, le cluster 1 regroupe la grande majorité des patients (94.1%), tandis que le cluster 2 représente un petit groupe, comprenant seulement 5.4% des patients.

Après avoir identifié les deux clusters distincts à partir de la CAH et confirmé leur présence à travers la visualisation du dendrogramme, nous nous intéressons à présent à la caractérisation de ces clusters et la comparaison des proportions des différentes variables entre eux-ci.

3.1.4 Interprétation des résultats

Identification des caractéristiques clés des clusters

Afin de caractériser les clusters, nous avons créé une variable qualitative nommée 'clusters' dont les modalités sont 'cluster 1' et 'cluster 2' (le cluster 1 regroupe le plus grand nombre de patients). Lorsqu'un patient appartient au cluster 1, cette variable prend la valeur 1 et 2 dans le cas contraire.

Afin d'étudier le lien entre une modalité de la variable 'clusters' et une modalité d'une autre variable qualitative du jeu de données, nous avons effectué un test exact de Fisher. Par exemple, si on considère la variable FL_HALLUCINOGENS, le test exact de Fisher peut permettre de tester si probabilité d'être dépendants aux hallucinogènes est égale dans les deux clusters.

Dans ce cas, les hypothèses sont les suivantes :

H0 : l n'y a aucune différence entre les proportions des patients dépendants aux hallucinogènes dans les clusters 1 et 2.

H1 : Il y a une différence entre les proportions des patients dépendants aux hallucinogènes dans les clusters 1 et 2. Plus précisément, il peut y avoir deux scénarios : soit la proportion de patients dépendants aux hallucinogènes est plus élevée dans le cluster 1 (la modalité "Oui" est sur-représentée dans le cluster 1), soit la proportion est plus faible dans le cluster 1 (la modalité "Oui" est sous-représentée dans le cluster 1).

la table de contingence peut être représentée comme suit (nous mettons des notations arbitraires afin de rendre la compréhension plus facile) :

		clusters		total
		cluster 1	cluster 2	
FL_HALLUCINOGENS	Oui	n_{11}	n_{12}	n_{m1}
	Non	n_{21}	n_{22}	n_{m2}
total		n_1	n_2	n

Sous H_0 , la variable aléatoire N_{m1} qui représente le nombre de patients possédant la modalité 'Oui' suit une loi hypergéométrique de paramètres n_1 , $\frac{n_{m1}}{n}$ et n . Si $P(N_{m1} \geq n_{m1})$ (c'est-à-dire la p-valeur du test) est inférieure à 0.05, alors la catégorie 'Oui' est sur représentée ou sous-représentée dans le cluster et permet ainsi d'affirmer que cette catégorie caractérise le cluster 1 de manière significative.

Cette p-valeur peut être transformée en quantiles d'une distribution normale centrée réduite appelée valeur test (notée v.test dans les tableaux 3.6 et 3.7). Si $v.test > 1.96$, la catégorie est sur-représentée dans le cluster. Sinon, elle y est sous-représentée. L'analyse des résultats contenus dans les tableaux 3.6 et 3.7 permet d'identifier les modalités des caractéristiques des patients qui caractérisent chaque cluster. Seules les modalités significatives y sont présentées. Les valeurs test y ont été classés de manière décroissante afin d'identifier les modalités qui caractérisent le plus chaque cluster.

caractérisation des clusters à l'aide des caractéristiques Socio-démographiques

Le cluster 1 représente les patients ayant déclaré vivre en couple, dans des domiciles fixes au cours des 30 derniers jours et dont le niveau d'instruction est le secondaire.

Le cluster 2 représente les patients dont les principaux revenus proviennent des allocations familiales, dont le niveau d'instruction est le primaire et qui ont déclaré avoir vécu avec des amis au cours du mois précédent.

caractérisation des clusters à l'aide des types de substances utilisées

Au regard des tableaux 3.6 et 3.7 on peut faire deux observations importantes :

Si on considère le cluster 1, on remarque que les valeurs test correspondant aux modalités 'Non' des variables liées aux types de substances sont toutes positives (Ces modalités

sont donc significativement sur-représentées dans le cluster 1. On note par ailleurs que ces modalités sont significativement sous-représentées dans le cluster 2).

Tandis que les valeurs test correspondant aux modalités 'Oui' des variables liées aux types de substances sont toutes positives dans le cluster 2 (Ces modalités sont donc significativement sur-représentées dans le cluster 2).

Ces résultats suggèrent que **le cluster 1 représente les patients ayant une dépendance aux substances moins sévère tandis que le cluster 2 représente ceux qui ont une dépendance aux substances plus sévère.**

caractérisation des clusters à l'aide du profil d'addiction et du référent

En ce qui concerne la variable CD_REFFERAL représentant la personne ou l'institution ayant incité le patient à suivre le traitement, la modalité 'justice' est sur-représentée parmi les patients appartenant au cluster 1 tandis que les modalités 'famille' et 'hôpital' sont sur-représentées dans le cluster 2. Au sujet de la fréquence de consommation du cannabis, la modalité 'Tous les jours' est sur-représentée dans le cluster 2 tandis qu'elle est sous-représentée dans le cluster 1.

L'utilisation des substances par injection permet de différencier davantage les clusters. Les patients ayant déclaré avoir déjà eu recours à une substance psychoactive par injection au cours de leur vie (variable CD_INJECTING_STAT= Oui) sont sur-représentés dans le cluster 2 tandis que ceux n'ayant jamais eu recours à une substance psychoactive par injection sont sur-représentés dans le cluster 1. Ce dernier résultat confirme davantage le fait que le cluster 2 représente les patients qui ont une addiction plus sévère aux substances.

TABLE 3.6 – modalités des variables qui caractérisent au mieux le cluster 1

	Cla/Mod(%) ^a	Mod/Cla(%) ^b	proportion globale(%) ^c	p.value ^d	v.test ^e
FL_HALLUCINOGENS=Non	97.320	99.714	96.968	0.000	17.108
FL_KETAMINE=Non	96.517	99.886	97.943	0.000	14.347
FL_HYPNOTICS=Non	96.561	99.600	97.618	0.000	13.825
FL_BENZODIAZEPINES=Non	96.198	99.886	98.267	0.000	12.925
FL_STIMULANTS=Non	96.804	95.309	93.178	0.000	11.004
FL_LSD=Non	95.519	100	99.080	0.000	9.859
FL_MDMA=Non	96.031	98.284	96.860	0.000	9.661
FL_COCAINE=Non	96.585	92.220	90.363	0.000	8.942
FL HEROIN=Non	95.259	100	99.350	0.000	8.177
FL_OPIATES=Non	95.353	99.771	99.025	0.000	7.963
FL_CRACK=Non	95.155	100	99.459	0.000	7.411
FL_MEPHEDRONE=Non	95.202	99.886	99.296	0.000	7.244
FL_GHB=Non	95.052	100	99.567	0.000	6.566
FL_AMPHETAMINES=Non	95.547	96.968	96.048	0.000	6.313
FL_HASH=Non	95.645	95.481	94.478	0.000	6.153
FL_POWDER_COCAINE=Non	95.582	95.309	94.369	0.000	5.786
FL_OTHER_HALLUCINOGENS=Non	94.997	99.943	99.567	0.000	5.769
CD_REFERRAL=Justice	96.680	41.648	40.769	0.001	3.294
CD_LVN_STAT_WHERE=Domicile fixe	95.284	90.160	89.551	0.001	3.238
FL_OTHER_OPIATES=Non	94.743	100	99.892	0.003	2.984
CD_EDUCATION_LEVEL=secondaire	95.801	63.959	63.184	0.004	2.843
CD_LVN_STAT_WITH_WHOM=En couple	97.492	17.792	17.271	0.008	2.644
FL_ALCOHOL=Non	95.164	87.815	87.331	0.014	2.449
CD_LVN_CHILDREN=Oui	96.857	19.394	18.950	0.034	2.123
FL_MARIJUANA=Non	95.924	43.078	42.501	0.034	2.116

ci-dessous, m désigne une modalité.

^a proportion de patients appartenant au cluster 1 parmi ceux qui possèdent la modalité m .

^b proportion de patients possédant la modalité m parmi ceux qui appartiennent au cluster 1.

^c proportion de patients possédant la modalité m dans l'ensemble des patients.

^d P-valeur du test exact de Fisher.

^e équivalent des p-valeurs en quantiles d'une distribution normale centrée réduite. Si $v.test > 1.96$, la modalité est sur-représentée dans le cluster 1. sinon, elle y est sous représentée.

TABLE 3.7 – modalités des variables qui caractérisent au mieux le cluster 2

	Cla/Mod(%) ^a	Mod/Cla(%) ^b	proportion globale(%) ^c	p.value ^e	v.test ^f
FL_HALLUCINOGENS=Oui	91.071	51.515	3.032	0.000	17.108
FL_KETAMINE=Oui	94.737	36.364	2.057	0.000	14.347
FL_HYPNOTICS=Oui	84.091	37.374	2.382	0.000	13.825
FL_BENZODIAZEPINES=Oui	93.750	30.303	1.733	0.000	12.925
FL_STIMULANTS=Oui	34.921	44.444	6.822	0.000	11.004
FL_LSD=Oui	100	17.172	0.920	0.000	9.859
FL_MDMA=Oui	48.276	28.283	3.140	0.000	9.661
FL_COCAINE=Oui	23.596	42.424	9.637	0.000	8.942
FL_HEROIN=Oui	100	12.121	0.650	0.000	8.177
FL_OPIATES=Oui	77.778	14.141	0.975	0.000	7.963
FL_CRACK=Oui	100	10.101	0.541	0.000	7.411
FL_MEPHEDRONE=Oui	84.615	11.111	0.704	0.000	7.244
FL_GHB=Oui	100	8.081	0.433	0.000	6.566
FL_AMPHETAMINES=Oui	27.397	20.202	3.952	0.000	6.313
FL_HASH=Oui	22.549	23.232	5.522	0.000	6.153
FL_POWDER_COCAINE=Oui	21.154	22.222	5.631	0.000	5.786
FL_OTHER_HALLUCINOGENS=Oui	87.500	7.071	0.433	0.000	5.769
CD_EDUCATION_LEVEL=primaire	8.108	42.424	28.045	0.002	3.145
FL_OTHER_OPIATES=Oui	100	2.020	0.108	0.003	2.984
CD_REFERRAL=Famille	10.169	18.182	9.583	0.007	2.718
CD_INCOME_STAT=Alloc.familiales	15.909	7.071	2.382	0.009	2.594
CD_REFERRAL=Hôpital	12.500	10.101	4.331	0.012	2.523
FL_ALCOHOL=Oui	8.974	21.212	12.669	0.014	2.449
CD_LVN_STAT_WITH_WHOM=Avec amis	11.538	9.091	4.223	0.028	2.192
CD_LABOUR_STAT=Autre	18.182	4.040	1.191	0.032	2.144
CD_LVN_CHILDREN=Non	5.878	88.889	81.050	0.034	2.123
FL_MARIJUANA=Oui	6.309	67.677	57.499	0.034	2.116
CD_FREQ_USE=Tous les jours	6.233	68.687	59.069	0.044	2.012

ci-dessous, m désigne une modalité.

^a proportion de patients appartenant au cluster 2 parmi ceux qui possèdent la modalité m .

^b proportion de patients possédant la modalité m parmi ceux qui appartiennent au cluster 2.

^c proportion de patients possédant la modalité m dans l'ensemble des patients.

^d P-valeur du test exact de Fisher.

^e Équivalent des p-valeurs en quantiles d'une distribution normale centrée réduite. Si $v.test > 1.96$, la modalité est sur-représentée dans le cluster 2. sinon, elle y est sous représentée.

En somme, l'analyse en cluster des patients ayant demandé pour la première fois en 2021 un traitement pour dépendance au cannabis nous a permis d'identifier 2 principaux clusters caractérisés comme suit :

- cluster 1 : patients ayant une dépendance aux substances moins sévère, envoyés en traitement par la justice, ayant déclaré vivre en couple, dans des domiciles fixes au cours des 30 derniers jours et dont le niveau d'instruction le plus élevé est le secondaire.
- cluster 2 : patients ayant une dépendance aux substances plus sévère, envoyés en traitement par les membres de la famille ou par un hôpital, dont les principaux revenus proviennent des allocations familiales ; le niveau d'instruction le plus élevé est le primaire et ces patients ont déclaré avoir vécu avec des amis au cours du mois précédent.

Comparaison entre les clusters

Dans toute la suite, les différences de proportion entre les clusters sont comparées à l'aide du test du chi carré. Les p-valeurs sont consignés dans les tableaux 3.8, 3.9 et 3.10.

Les différences de moyennes d'âges entre les clusters sont comparées à l'aide du test de student.

Commençons par comparer les clusters sur le rapport des caractéristiques socio-démographiques.

Comparaison entre les caractéristiques Socio-démographiques au sein des deux clusters

Comme le montre le tableau 3.8, il y'a des différences significatives entre les clusters en ce qui concerne l'âge lors du premier usage du cannabis, le niveau d'instruction, la situation du logement et le type de logement.

Nous n'avons pas trouvé de différence significative entre l'âge des patients au sein des deux clusters ($p = 0.19$). Les moyennes d'âges sont relativement similaires (environ 25 ans dans le cluster 1 et 24 ans dans le cluster 2). Par contre, on note une différence significative dans l'âge lors du premier usage du cannabis entre les deux clusters $p(< 0.05)$: les patients du cluster 2 ont fait usage du cannabis pour la première fois relativement plus tôt que ceux du cluster 1 (l'âge lors du premier usage du cannabis au sein du cluster 2 est significativement inférieure à celui du cluster 1).

Par rapport au cluster 1, une proportion significativement plus élevée de patients appartenant au cluster 2 a un niveau d'instruction primaire, vit avec un parent et dans des domiciles fixes. Par rapport au cluster 2, une proportion significativement plus élevée de patients du cluster 1 ont le niveau secondaire et vivent avec un parent et dans des domiciles fixes.

TABLE 3.8 – Différences entre les caractéristiques socio-démographiques des patients au sein des clusters

		tests statistiques ^a			
		Cluster 1 (N=1748) %(n)	Cluster 2 (N=99) %(n)	t	test χ^2
NR__AGE	Moyenne(écart-type)	25.93(9.10)	24.73(7.84)	0.19	
	Range	13-64	15-45		
MS__AGE__FIRST__USE	Moyenne(écart-type)	15.93(0.08)	15.07(0.27)	0.00	
	Range	7-51	10-30		
CD__SEX					1
	Hommes	80.09(1400)	79.8(79)		
	Femmes	19.90(348)	20.20(20)		
CD__INCOME__STAT					0.054
	salaire	30.26()	25.25()		
	alloc.chômage	9.55(167)	13.13(13)		
	pension	0.23(4)	0		
	indemnités	13.15(230)	8.08(8)		
	CPAS	7.66(134)	6.06(6)		
	alloc.familiales	2.11(37)	7.07(7)		
	bourse d'études	0.11(2)	0		
	aucun revenu	36.27(634)	40.40(40)		
	autre	0.63(11)	0		
CD__EDUCATION__LEVEL					0.02
	aucun	1.83(32)	2.02(2)		
	primaire	27.23(476)	42.42(42)		
	secondaire	63.95(1118)	49.49(49)		
	supérieure	6.06(106)	5.05(5)		
	autre	0.91(16)	1.01(1)		
CD__LVN__STAT__WITH__WHOM					0.01
	seul	25.91(453)	27.27(27)		
	en couple	17.79(311)	8.08(8)		
	avec parent(s)	46.45(812)	47.47(47)		
	avec famille	4.97(87)	8.08(8)		
	avec amis	3.94(69)	9.09(9)		
	autre	0.91(16)	0		
CD__LABOUR__STAT					0.15
	régulier	28.71(502)	24.24(24)		
	occasionnel	2.63(46)	2.02(2)		
	chômage	17.84(312)	22.22(22)		
	aux études	36.27(634)	37.37(37)		
	en incapacité de travail	12.18(213)	10.10(10)		
	au foyer	1.20(21)	0		
	pensionné/pré-pensionné	0.11(2)	0		
	autre	1.03(18)	4.04(4)		
CD__LVN__STAT__WHERE					0.00
	Autre	0.57(10)	0		
	Domicile fixe	90.16(1576)	78.78(78)		
	Institution	3.71(65)	8.08(8)		
	Logements variables	3.60(63)	6.06(6)		
	Prison	1.20(21)	4.04(4)		
	Rue	0.74(13)	3.03(3)		
CD__LVN__CHILDREN					0.055
	oui	19.39(339)	80.60(11)		
	non	80.60(1409)	88.88(88)		

^a Seules les p-valeurs sont présentées dans ce tableau. Les valeurs en gras expriment la significativité au seuil de 5%.

Comparaison entre les types de substances utilisées au sein des deux clus-

ters

Comme le montre le tableau 3.9, il y a des différences significatives entre les clusters en ce qui concerne la dépendance aux stimulants (cocaïne ou autres), aux opiacés (autres opiacés), aux hypnotiques ou sédatifs (benzodiazepines, GHB), aux hallucinogènes (LSD (Lysergide), kétamine ou autres), à l'alcool.

La consommation des substances problématiques autres que le cannabis est plutôt peu fréquente au sein du cluster 1. Par exemple, ce dernier se distingue par l'absence de tout patient ayant déclaré avoir eu un problème avec l'héroïne, le crack, le GHB, le LSD, les hallucinogènes (non spécifiés) et les opiacés (non spécifiés).

Comparativement au cluster 1, significativement plus de patients appartenant au cluster 2 ont déclaré avoir comme substance problématique de la cocaïne en poudre (22.2% contre 4.69%), aux amphétamines (20.20% contre 3.03%), à la MDMA (71.71% contre 1.71%), à la méphédronne (11.11% contre 0.11%), les benzodiazepines (30.30% contre 0.11%), la kétamine (36.36% contre 1.71%) et l'alcool (21.21% contre 12.18%).

TABLE 3.9 – Différences entre les types de substances utilisées au sein des clusters

		Cluster 1 (N=1748) %(n)	Cluster 2 (N=99) % (n)	test χ^2
Opiacés	FL_HEROIN	0	12.12(12)	0.00^a
	FL_METHADONE	0	0	_b
	FL_BUPRENORPHINE	0	0	-
	FL_FENTANYL	0	0	-
	FL_OTHER_OPIATES	0	2.02(2)	0.00
Cocaïne	FL_POWDER_COCAINE	4.69(82)	22.22(22)	0.00
	FL_CRACK	0	10.10(89)	0.00
	FL_OTHER_COCAINE	0	1.01(1)	0.04
Stimulants (autres que cocaïne)	FL_AMPHETAMINES	3.03(53)	20.20(20)	0.00
	FL_METHAMPHETAMINES	0.11(2)	1.01(1)	0.38
	FL_MDMA	1.71(30)	71.71(28)	0.00
	FL_MEPHEDRONE	0.11(2)	11.11(11)	0.00
	FL_OTHER_STIMULANTS	0.05(1)	0	1
Hypnotiques ou sédatifs	FL_BARBITURATES	0	0	-
	FL_BENZODIAZEPINES	0.11(2)	30.30(30)	0.00
	FL_GHB	0	8.08(8)	0.00
	FL_OTHER_HYPNOTICS	0	0	-
Hallucinogènes	FL_LSD	0	17.17(17)	0.00
	FL_KETAMINE	0.11(2)	36.36(36)	0.00
	FL_OTHER_HALLUCINOGENS	0.05(1)	7.07(7)	0.00
Alcool	FL_ALCOHOL	12.18(213)	21.21(21)	0.01
Inhalants volatils	FL_VOLATILE	0.22(4)	0	1
Autres substances	FL_OTHER_SUBSTANCE	0	0	-

^a Seules les p-valeurs sont présentées dans ce tableau. Les valeurs en gras expriment la significativité au seuil de 5%.

^b Test non concluant car trop peu d'observations. Ces variables sont incluses dans ce tableau juste pour visualiser les proportions au sein des clusters.

La fréquence avec laquelle les patients avaient consommé le cannabis au cours des 30 derniers jours a été incluse dans les analyses mais n'est pas présentée dans ce tableau car la plupart ont déclaré avoir consommé le cannabis tous les jours.

Comparaison entre le profil d'addiction et l'origine du renvoi en traitement au sein des deux clusters

Il y a des différences significatives dans l'origine du renvoi en traitement entre les deux clusters. Une proportion significativement plus élevée de patients au sein du cluster 1 ont été invités ou contraints à suivre le traitement par la justice (41.61% contre 25.25%).

TABLE 3.10 – Différences entre la fréquence d'utilisation du cannabis et l'origine du renvoi en traitement au sein des clusters

		Cluster 1 (N=1748) %(n)	Cluster 2 (N=99) %(n)	test χ^2
CD_REFERRAL				0.00^a
	Individuel	24.08(421)	30.30(30)	
	Amis	2.40(42)	1.01(1)	
	Famille	9.09(159)	18.18(18)	
	Centres pour toxicomanes	1.08(19)	0	
	Justice	41.64(728)	25.25(25)	
	Hôpital	4(70)	10.10(10)	
	Médecin généraliste	4.91(86)	1.01(1)	
	Autre service médical	8.92(156)	11.11(11)	
	Autre	3.83(67)	3.03(3)	

^a P-valeur du test χ^2 .

En somme, les résultats du clustering appliqué aux données des patients entrés en traitement pour une addiction au cannabis ont abouti à l'identification de deux clusters optimaux distincts. Ces clusters ont révélé des différences significatives au niveau des caractéristiques socio-démographiques, du profil d'addiction et de l'origine du renvoi en traitement entre les clusters. Ces observations mettent en évidence la diversité des profils parmi les patients entrant en traitement pour usage du cannabis. Afin d'explorer davantage ces profils de patients, nous allons maintenant examiner l'évolution des profils des patients au fil du temps, sur une période de cinq ans allant de 2015 à 2020. Cette analyse nous permettra de déterminer si les profils de patients observés sur cette période sont similaires ou non à ceux observés sur l'année 2021.

3.2 Évolution des profils des patients au cours du temps : Comparaison des profils des patients entrés en traitement pour la première fois entre 2015 et 2020 avec ceux de l'année 2021

Dans cette section, nous analysons les similitudes et les différences entre les clusters de patients ayant demandé un traitement pour la première fois au cours de la période allant de 2015 à 2020 avec ceux de l'année 2021. Le but est d'identifier les tendances et les changements dans les profils de ces patients au fil du temps.

Dans les sections précédentes, nous avons identifié les clusters pour l'année 2021 à l'aide de la CAH. Dans un souci de cohérence dans les analyses, nous choisissons d'utili-

ser également la CAH pour identifier les clusters des patients pour les années comprises entre 2015 et 2020. Ainsi, nous pourrions explorer les regroupements obtenus sur chaque année et analyser les similitudes et les différences avec ceux obtenus en 2021. Tout comme sur le jeu de donnée de 2021, nous avons appliqué une AFDM dans le but de traiter les données d'entrée qui sont mixtes et capturer la variance totale. Sur les données de chaque année, cette analyse nous a permis de générer de nouvelles variables continues à partir des données initiales, qui ont été utilisées comme variables d'entrée pour l'application de la CAH.

3.2.1 Évolution du nombre de patients enregistrés pour la première fois entre 2015 et 2021, pour un traitement lié à l'usage du cannabis

Comme le montre la figure 3.11, le nombre de demandes de traitement pour le cannabis est quasi stable dans le temps, à l'exception des années 2020 et 2021 où on note une baisse significative.

Cette baisse des demandes de traitement observée entre 2020 et 2021 peut être attribuée à la pandémie de COVID-19. En effet, la crise sanitaire a non seulement perturbé l'offre de services de traitement aux consommateurs de cannabis, mais elle a également affecté le rapportage précis et complet des demandes de traitement par les centres de traitement à Sciensano [Antoine, 2021]. Ainsi, il est possible que la diminution observée reflète en partie les conséquences de la pandémie sur la disponibilité des services et sur la capacité des usagers de cannabis à accéder aux traitements appropriés.

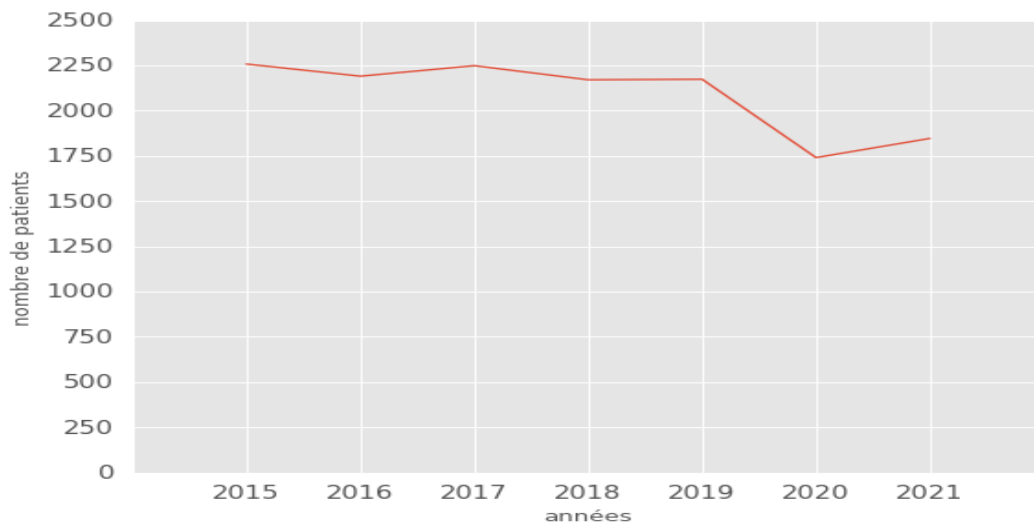


FIGURE 3.11 – Évolution du nombre de patients enregistrés pour la première fois entre 2015 et 2021, pour un traitement lié à l'usage du cannabis

Après avoir examiné l'évolution du nombre de demandes de traitement enregistrées au fil du temps, nous allons à présent comparer les profils observés sur la période allant de 2015 à 2020 avec ceux de l'année 2021.

3.2.2 Évolution des profils des patients au cours du temps : comparaison des profils des patients entrés en traitement pour la première fois entre 2015 et 2020 avec ceux de l'année 2021

L'analyse en clusters par année des patients traités pour une addiction au cannabis entre 2015 et 2020 a montré des profils similaires à ceux observés en 2021, à l'exception de l'année 2019 où une différence significative a été constatée (Pour chaque année, les scores de silhouette moyen en fonction du nombre de clusters ainsi que les dendrogrammes correspondant se trouvent en annexe A). En effet, lors de cette année spécifique, trois clusters optimaux distincts ont été identifiés, ce qui indique une variation dans les profils des patients par rapport aux autres années étudiées.

Étant donné l'observation de profils différents en 2019 par rapport aux autres années, une attention particulière sera accordée à cette année dans la suite. Nous allons examiner de plus près les caractéristiques spécifiques des patients inclus dans les trois clusters identifiés, afin de comprendre les facteurs sous-jacents qui ont conduit à cette différence par rapport à l'année 2021. Toutefois, il est important de souligner que ce troisième cluster identifié est extrêmement limité, ne comprenant qu'un total de 3 patients.

Comme le montre la figure 3.12, ces clusters identifiés présentent des proportions variées : le cluster 1 représente la grande majorité avec 95.1% des patients, indiquant une prévalence élevée de ce profil parmi les patients ayant demandé un traitement pour la première fois en 2019. Le cluster 2, quant à lui, comprend 4.7% des patients, représentant un groupe de patients plus restreint mais significatif. Enfin, le cluster 3 est le moins fréquent, ne comprenant que 0.1% des patients.

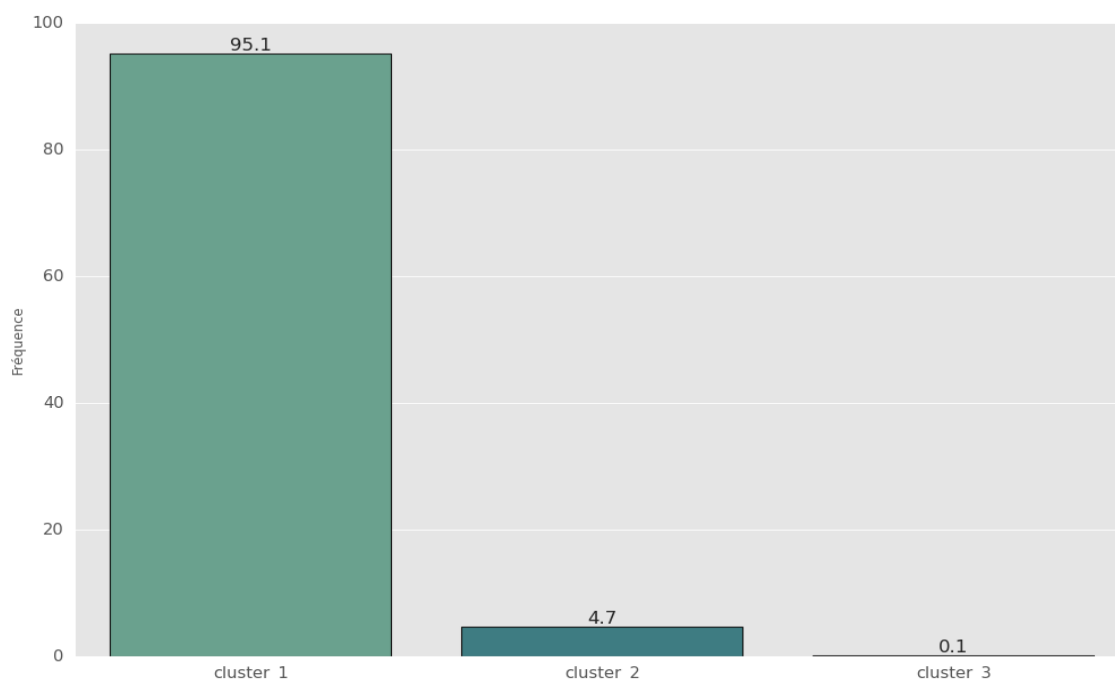


FIGURE 3.12 – Répartition des patients au sein des 3 clusters : année 2019

Les tableaux 3.11, 3.12 et 3.13 présentent les modalités des variables significativement sur-représentées au sein de chaque cluster.

On remarque que les deux premiers clusters présentent des similarités marquantes avec ceux observés en 2021, même si on peut relever de légères différences. Le premier cluster représente le plus grand groupe de patients, caractérisé par une dépendance aux substances moins sévère. Ces patients sont principalement envoyés en traitement par la justice, ont un travail régulier et ont un niveau d'instruction secondaire. En comparaison, le deuxième cluster comprend des patients présentant une dépendance plus sévère aux substances. Ce groupe est principalement composé d'étudiants ou de patients en formation, ayant un niveau d'instruction primaire. De plus, ces patients ont une tendance plus fréquente à recourir à l'injection de substances.

Enfin, le troisième cluster se distingue des deux premiers par sa composition principalement constituée de patients pré-pensionnés ou pensionnés. Ces patients ont également fréquemment recours à l'injection de substances.

TABLE 3.11 – modalités des variables qui caractérisent au mieux le cluster 1

	Cla/Mod(%) ^a	Mod/Cla(%) ^b	proportion globale(%) ^c	p.value ^e	v.test ^f
FL_HALLUCINOGENS=Non	98.140	99.516	96.458	0.000	19.906
FL_KETAMINE=Non	97.402	99.710	97.378	0.000	17.189
FL_HYPNOTICS=Non	96.582	99.758	98.252	0.000	13.360
FL_BENZODIAZEPINES=Non	96.359	99.807	98.528	0.000	12.267
FL_LSD=Non	96.186	100	98.896	0.000	12.039
FL_STIMULANTS=Non	97.088	95.116	93.192	0.000	11.305
FL_MDMA=Non	96.498	98.598	97.194	0.000	10.961
FL_OPIATES=Non	95.909	99.758	98.942	0.000	9.288
FL_OTHER_HALLUCINOGENS=Non	95.739	99.952	99.310	0.000	8.726
FL_HEROIN=Non	95.652	100	99.448	0.000	8.308
FL_COCAINE=Non	96.571	92.602	91.214	0.000	7.991
FL_POWDER_COCAINE=Non	96.203	95.551	94.480	0.000	7.293
FL_AMPHETAMINES=Non	96.013	96.663	95.768	0.000	6.770
FL_METHAMPHETAMINES=Non	95.431	100	99.678	0.000	6.207
FL_GHB=Non	95.343	100	99.770	0.000	5.157
FL_OTHER_OPIATES=Non	95.343	100	99.770	0.000	5.157
FL_ALCOHOL=Non	96.046	88.104	87.259	0.000	4.622
FL_VOLATILE=Non	95.289	99.758	99.586	0.001	3.439
CD_INJECTING_STAT=Non	95.300	99.033	98.850	0.007	2.694
FL_HASH=Non	95.538	91.103	90.708	0.010	2.560
CD_EDUCATION_LEVEL=secondaire	96.009	62.814	62.236	0.016	2.418
CD_REFERRAL=Justice	96.370	37.234	36.753	0.038	2.078
CD_FREQ_USE=Pas le mois précédent	97.473	13.056	12.741	0.042	2.038
CD_LABOUR_STAT=Régulier	96.616	27.611	27.185	0.044	2.014
FL_FENTANYL=Non	95.168	100	99.954	0.049	1.971
FL_METHADONE=Non	95.168	100	99.954	0.049	1.971
FL_MARIJUANA=Non	96.090	47.534	47.056	0.049	1.970

ci-dessous, m désigne une modalité.

^a proportion de patients appartenant au cluster 1 parmi ceux qui possèdent la modalité m .

^b proportion de patients possédant la modalité m parmi ceux qui appartiennent au cluster 1.

^c proportion de patients possédant la modalité m dans l'ensemble des patients.

^d proportion de patients possédant la modalité m dans l'ensemble des patients.

^e P-valeur du test exact de Fisher.

^f Équivalent des p-valeurs en quantiles d'une distribution normale centrée réduite. Si $v.test > 1.96$, la modalité est sur-représentée dans le cluster. sinon, elle y est sous représentée.

TABLE 3.12 – modalités des variables qui caractérisent au mieux le cluster 2

	Cla/Mod(%) ^a	Mod/Cla(%) ^b	proportion globale(%) ^c	p.value ^e	v.test ^f
FL_HALLUCINOGENS=Oui	87.013	65.049	3.542	0.000	20.057
FL_KETAMINE=Oui	89.474	49.515	2.622	0.000	17.304
FL_HYPNOTICS=Oui	86.842	32.039	1.748	0.000	13.444
FL_BENZODIAZEPINES=Oui	87.500	27.184	1.472	0.000	12.342
FL_LSD=Oui	100	23.301	1.104	0.000	12.103
FL_STIMULANTS=Oui	31.757	45.631	6.808	0.000	11.448
FL_MDMA=Oui	52.459	31.068	2.806	0.000	11.056
FL_OPIATES=Oui	78.261	17.476	1.058	0.000	9.347
FL_OTHER_HALLUCINOGENS=Oui	93.333	13.592	0.690	0.000	8.774
FL_HEROIN=Oui	100	11.650	0.552	0.000	8.351
FL_COCAINE=Oui	19.895	36.893	8.786	0.000	8.129
FL_POWDER_COCAINE=Oui	23.333	27.184	5.520	0.000	7.401
FL_AMPHETAMINES=Oui	25	22.330	4.232	0.000	6.863
FL_METHAMPHETAMINES=Oui	100	6.796	0.322	0.000	6.239
FL_GHB=Oui	100	4.854	0.230	0.000	5.184
FL_OTHER_OPIATES=Oui	100	4.854	0.230	0.000	5.184
FL_ALCOHOL=Oui	11.191	30.097	12.741	0.000	4.765
FL_VOLATILE=Oui	44.444	3.883	0.414	0.001	3.469
CD_EDUCATION_LEVEL=primaire	7.187	45.631	30.083	0.001	3.396
FL_HASH=Oui	8.911	17.476	9.292	0.008	2.669
CD_LABOUR_STAT=Aux études	6.250	47.573	36.063	0.014	2.445
CD_INJECTING_STAT=Oui	16	3.883	1.150	0.033	2.129
FL_MARIJUANA=Oui	5.647	63.107	52.944	0.034	2.118
CD_LVN_STAT_WITH_WHOM=Avec Famille	9.259	9.709	4.968	0.040	2.050
FL_FENTANYL=Oui	100	0.971	0.046	0.047	1.983
FL_METHADONE=Oui	100	0.971	0.046	0.047	1.983

ci-dessous, m désigne une modalité.

^a proportion de patients appartenant au cluster 2 parmi ceux qui possèdent la modalité m .

^b proportion de patients possédant la modalité m parmi ceux qui appartiennent au cluster 2.

^c proportion de patients possédant la modalité m dans l'ensemble des patients.

^d proportion de patients possédant la modalité m dans l'ensemble des patients.

^e P-valeur du test exact de Fisher.

^f Équivalent des p-valeurs en quantiles d'une distribution normale centrée réduite. Si $v.test > 1.96$, la modalité est sur-représentée dans le cluster. sinon, elle y est sous représentée.

TABLE 3.13 – modalités des variables qui caractérisent au mieux le cluster 3

	Cla/Mod(%) ^a	Mod/Cla(%) ^b	proportion globale(%) ^c	p.value ^e	v.test ^f
CD_INCOME_STAT=Pension	100	100	0.138	0.000	6.194
CD_LABOUR_STAT=Pensionné/pré-pensionné	100	100	0.138	0.000	6.194
CD_INJECTING_STAT=Oui	4	33.333	1.150	0.034	2.114

ci-dessous, m désigne une modalité.

^a proportion de patients appartenant au cluster 3 parmi ceux qui possèdent la modalité m .

^b proportion de patients possédant la modalité m parmi ceux qui appartiennent au cluster 3.

^c proportion de patients possédant la modalité m dans l'ensemble des patients.

^d proportion de patients possédant la modalité m dans l'ensemble des patients.

^e P-valeur du test exact de Fisher.

^f Équivalent des p-valeurs en quantiles d'une distribution normale centrée réduite. Si $v.test > 1.96$, la modalité est sur-représentée dans le cluster. sinon, elle y est sous représentée.

En poursuivant notre analyse, nous allons maintenant examiner de manière approfondie les caractéristiques socio-démographiques, les types de substances utilisées et les profils d'addiction au sein de chaque cluster. Cette comparaison nous permettra de mettre en évidence d'éventuelles différences significatives entre les clusters.

Comparaison des caractéristiques socio-démographiques des patients au sein des clusters identifiés en 2019

L'analyse comparative des clusters (tableau 3.14) met en évidence des différences significatives en termes d'âge, d'âge lors du premier usage du cannabis, de source de revenus, de niveau d'instruction et de la situation du travail.

Le cluster 1 est composé principalement de jeunes adultes, avec une moyenne d'âge de 26 ans. Le cluster 2 présente également une population jeune, avec une moyenne d'âge légèrement inférieure de 23.3 ans. En revanche, le cluster 3 se démarque par une moyenne d'âge beaucoup plus élevée de 57.3 ans.

Concernant l'âge lors du premier usage du cannabis, le Cluster 1 et le Cluster 2 ont des âges moyens similaires, respectivement 16 ans et 15.5 ans. En revanche, le Cluster 3 se distingue par un âge moyen beaucoup plus élevé de 26 ans.

On note également des différences significatives en termes de répartition des revenus entre les clusters : le cluster 1 se distingue par une proportion plus élevée de patients ayant un salaire, des allocations chômage, des indemnités, des allocations familiales ou aucune source de revenu. Le cluster 2 est caractérisé par une proportion plus élevée de patients encore aux études, bénéficiant d'une bourse d'études. Enfin, le cluster 3 se compose principalement de patients pré-pensionnés ou pensionnés.

Comme sur l'année 2021, le cluster 1 compte une proportion plus élevée de patients ayant un niveau d'instruction secondaire, tandis que le cluster 2 se distingue par une proportion plus élevée de patients ayant un niveau d'éducation primaire. Le cluster 3 est caractérisé par une proportion plus élevée de patients ayant un niveau d'éducation supérieur.

Le cluster 1 regroupe une proportion plus élevée de personnes ayant un emploi régulier ou étant aux études. Le cluster 2 se caractérise par une proportion légèrement plus élevée de personnes au chômage. Le cluster 3 est principalement composé de personnes pré-pensionnés ou pensionnés.

TABLE 3.14 – Différences entre les caractéristiques socio-démographiques des patients au sein des clusters : 2019

		tests statistiques ^a			
		Cluster 1 (N=2068) %(n)	Cluster 2 (N=103) %(n)	Cluster 3 (N=3) %(n)	F test χ^2
NR_AGE	Moyenne(écart-type)	26(9.16)	23.3(8.23)	57.3(11)	0.00
	Range	13-64	14-49	45-66	
MS_AGE_FIRST_USE	Moyenne(écart-type)	16.0(3.68)	15.5(3.27)	26(18.2)	0.00
	Range	6-53	10-31	15-47	
CD_SEX					0.55
	Hommes	81.96(1695)	78.64(22)	66.67(2)	
	Femmes	18.04(373)	21.36(22)	33.33(1)	
CD_INCOME_STAT					0.00
	salaire	28.14(582)	25.24(26)	0	
	alloc.chômage	11.12(230)	11.65(12)	0	
	pension	0.43(9)	0	100(3)	
	indemnités	12.57(260)	8.74(9)	0	
	CPAS	8.66(179)	9.71(10)	0	
	alloc.familiales	4.45(92)	4.85(5)	0	
	bourse d'études	0	2.91(3)	0	
	aucun revenu	33.75(698)	38.83(40)	0	
	autre	0.68(14)	0	0	
CD_EDUCATION_LEVEL					0.01
	aucun	2.08(43)	0	0	
	primaire	29.35(607)	45.63(47)	0	
	secondaire	62.81(1299)	50.49(47)	0	
	supérieure	5.37(111)	3.88(4)	33.33(1)	
	autre	0.39(8)	0	0	
CD_LVN_STAT_WITH_WHOM					0.137
	seul	23.11(478)	21.36(22)	66.67(2)	
	en couple	18.04(373)	15.53(16)	33.33(1)	
	avec parent(s)	50.34(1041)	46.60(48)	0	
	avec famille	4.74(98)	9.71(10)	0	
	avec amis	3.14(65)	6.80(7)	0	
	autre	0.63(13)	0	0	
CD_LABOUR_STAT					0.00
	régulier	27.61(571)	19.42(20)	0	
	occasionnel	2.56(53)	3.88(4)	0	
	chômage	19.25(398)	21.36(22)	0	
	aux études	35.54(7.35)	37.37(49)	0	
	en incapacité de travail	11.46(237)	5.83(6)	0	
	au foyer	2.32(48)	0.97(1)	0	
	pensionné/pré-pensionné	0	0	100(3)	
	autre	1.26(26)	0.97(1)	0	
CD_LVN_STAT_WHERE					0.84
	Autre	0.39(8)	0	0	
	Domicile fixe	88.39(1828)	84.47(87)	100(3)	
	Institution	3.92(81)	3.88(4)	0	
	Logements variables	3.97(82)	6.80(7)	0	
	Prison	1.60(33)	3.88(1)	0	
	Rue	1.74(36)	3.88(4)	0	
CD_LVN_CHILDREN					0.19
	oui	15.43(319)	9.71(10)	33.33(1)	

^a Seules les p-valeurs sont présentées dans ce tableau. Les valeurs en gras expriment la significativité au seuil de 5%.

Comparaison du type de substances utilisées par les patients au sein des clusters identifiés en 2019

comme le montre le tableau 3.15, on note une différence significative en terme de types de substances utilisées entre les clusters ; Le Cluster 1 se caractérise par une absence d'uti-

lisation d'opiacés, mais une utilisation relativement élevée de cocaïne en poudre (4.45%), d'amphétamines (3.34%) et de benzodiazépines (0.19%). En revanche, le Cluster 2 présente une prévalence plus élevée d'opiacés, notamment l'héroïne (11.65%) et le fentanyl (0.97%). De plus, on observe une utilisation importante d'amphétamines (22.33%) et de benzodiazépines (27.18%) dans ce cluster. Le Cluster 3 se distingue par une utilisation moins fréquente de substances, avec des pourcentages très faibles dans toutes les catégories, à l'exception notable de l'utilisation de LSD (23.3%) par rapport aux hallucinogènes. Il convient également de souligner l'absence de consommation d'alcool dans le Cluster 3.

TABLE 3.15 – Différences entre les types de substances utilisées au sein des clusters : 2019

		Cluster 1 (N=2068) %(n)	Cluster 2 (N=103) %(n)	Cluster 3 (N=3) %(n)	test χ^2
Opiacés	FL_HEROIN	0	11.65(12)	0	0.00^a
	FL_METHADONE	0	0.97(1)	0	^b
	FL_BUPRENORPHINE	0	0	0	-
	FL_FENTANYL	0	0.97(1)	0	-
	FL_OTHER_OPIATES	0	4.85(5)	0	0.00
Cocaïne	FL_POWDER_COCAINE	4.45(92)	27.18(28)	0	0.00
	FL_CRACK	0.53(11)	0	0	0.75
	FL_OTHER_COCAINE	0	0	0	-
Stimulants (autres que cocaïne)	FL_AMPHETAMINES	3.34(69)	22.33(23)	0	0.00
	FL_METHAMPHETAMINES	0	6.8(7)	0	0.00
	FL_MDMA	1.40(29)	31.07(32)	0	0.00
	FL_MEPHEDRONE	0.39(8)	0.97(1)	0	0.66
	FL_OTHER_STIMULANTS	0	0	0	-
Hypnotiques ou sédatifs	FL_BARBITURATES	0	0	0	-
	FL_BENZODIAZEPINES	0.19(4)	27.18(28)	0	0.00
	FL_GHB	0	4.85(5)	0	0.00
	FL_OTHER_HYPNOTICS	0	0	0	-
Hallucinogènes	FL_LSD	0	23.3(24)	0	0.00
	FL_KETAMINE	0.29(6)	49.51(51)	0	0.00
	FL_OTHER_HALLUCINOGENS	0.05(1)	13.59(14)	0	0.00
Alcool					
Inhalants volatils	FL_ALCOHOL	11.19(246)	30.1(31)	0	0.00
	FL_VOLATILE	0.24(5)	3.88(4)	0	0.00
Autres substances					
	FL_OTHER_SUBSTANCE	0	0	0	-

^a Seules les p-valeurs sont présentées dans ce tableau. Les valeurs en gras expriment la significativité au seuil de 5%.

^b Test non concluant car trop peu de cellules. Ces variables sont incluses dans ce tableau juste pour visualiser les proportions au sein des clusters.

Note : La fréquence avec laquelle les patients avaient consommé le cannabis au cours des 30 derniers jours a été incluse dans les analyses mais n'est présentée dans ce tableau car la plupart ont déclaré avoir consommé le cannabis tous les jours.

Comparaison du profil d'addiction des patients entre les clusters identifiés en 2019

Comme le montre le tableau 3.16, on observe une différence significative de proportions en termes de profil d'addiction entre les clusters.

Dans le Cluster 1, la plupart des patients ont été référés au traitement par la justice

(37.23%), tandis que dans le Cluster 2, cette proportion est plus faible (26.21%). En revanche, dans le Cluster 3, l'orientation par la justice est plus élevée (66.67%), bien que l'échantillon soit plus restreint. Les renvois en traitement par la famille sont similaires dans les Clusters 1 (12.14%) et 2 (11.65%), mais elles sont absentes dans le Cluster 3. En ce qui concerne la fréquence d'utilisation du cannabis, le Cluster 1 présente une proportion plus élevée de patients consommant du cannabis quotidiennement (23.84%), tandis que dans le Cluster 2, cette proportion est légèrement inférieure (27.18%). Enfin, les Cluster 2 et 3 se distinguent par une proportion significativement plus élevée de patients ayant recours à l'injection de substances (3.88% et 33.33% respectivement) par rapport au cluster 1.

TABLE 3.16 – Différences entre la fréquence d'utilisation du cannabis et l'origine du renvoi en traitement au sein des clusters : 2019

		Cluster 1 (N=2068) %(n)	Cluster 2 (N=103) %(n)	Cluster 3 (N=3) %(n)	test χ^2
CD_REFERRAL					0.03^a
	Individuel	23.84(493)	27.18(28)	0	
	Amis	2.03(42)	1.94(2)	0	
	Famille	12.14(251)	11.65(12)	33.33(1)	
	Centres pour toxicomanes	2.32(48)	1.94(2)	0	
	Justice	37.23(770)	26.21(27)	66.67(2)	
	Hôpital	4.74(98)	8.74(9)	0	
	Médecin généraliste	5.13(106)	4.85(5)	0	
	Autre service médical	9.67(200)	12.62(13)	0	
	Autre	2.90(60)	4.85(5)	0	
CD_INJECTING_STAT		0.97(20)	3.88(4)	33.33(1)	0.00

^a P-valeur du test χ^2 .

En somme, l'analyse en cluster des patients entrés en traitement pour la première fois pour sur la période 2015 à 2020, a permis d'identifier deux clusters principaux avec des caractéristiques distinctes, similaires à ceux observés en 2021. Le premier cluster regroupe des patients présentant une dépendance aux substances moins sévère, référés en traitement principalement par la justice, vivant en couple, dans des domiciles fixes, et ayant un niveau d'instruction secondaire. Le deuxième cluster concerne les patients ayant une dépendance plus sévère, référés en traitement par les membres de la famille ou par un hôpital, avec des revenus principalement issus des allocations familiales, un niveau d'instruction primaire, et ayant déclaré avoir vécu avec des amis récemment. Ces observations soulignent la stabilité des profils de dépendance au cannabis au fil des années, avec ces deux clusters prédominants. Cependant, une analyse plus approfondie de l'année 2019 a révélé l'existence d'un troisième cluster distinct, principalement composé de patients pré-pensionnés ou pensionnés, qui ont également une propension élevée à recourir à l'injection de substances.

Chapitre 4

Discussion/Conclusion

Dans le cadre cette étude, notre objectif était d’effectuer une analyse en cluster afin de caractériser les profils des personnes ayant entamé un traitement pour la première fois en raison de leur dépendance au cannabis, sur la période allant de 2015 à 2021. Le type de substances consommées (opiacés, cocaïne, hallucinogènes, alcool, inhalants volatiles, stimulants autres que la cocaïne...), le profil d’addiction (fréquence d’utilisation du cannabis, origine du renvoi en traitement) et les caractéristiques socio-démographiques (sexe, âge, niveau d’instruction, situation de logement, etc.) ont été utilisées afin d’effectuer le regroupement.

Dans cette étude, nous avons dans un premier temps concentré nos analyses sur l’année 2021, en raison du contexte singulier qui l’entoure, caractérisé par un assouplissement progressif des mesures restrictives liées à la pandémie de COVID-19 ; le but visé étant une évaluation de l’influence potentielle de ces mesures sur les profils des patients. En analysant les données spécifiques à l’année 2021, nous avons pu obtenir des informations récentes et pertinentes sur les profils d’addiction au cannabis observés dans un contexte post-pandémique. Une fois ces résultats obtenus, nous avons ensuite élargi notre analyse pour y inclure la période allant de 2015 à 2020. Cette approche nous a permis d’évaluer la stabilité des profils des patients au fil du temps. En comparant les profils identifiés en 2021 avec ceux observés sur la période antérieure, nous avons pu déterminer si les mêmes tendances et caractéristiques se maintiennent ou s’ils ont évolué au fil des années. Cette approche temporelle nous a offert une perspective plus complète et nous a permis de vérifier la robustesse des clusters identifiés, en nous assurant qu’ils ne sont pas spécifiques à une année particulière.

Pour identifier les différents clusters de patients en 2021, nous nous sommes appuyés sur une comparaison entre différentes techniques de clustering (K-means, K-modes, K-prototypes ainsi que la classification ascendante hiérarchique). En privilégiant celle offrant le score de silhouette le plus élevé (cette métrique mesure à la fois la cohérence interne des clusters et leur séparation les uns des autres), nous avons maximisé la qualité de la séparation obtenue. Toutes ces techniques ont convergé vers l’identification de deux clusters optimaux, à l’exception de K-modes qui privilégiait une séparation en 3 clusters. La classification ascendante hiérarchique s’étant démarquée par le score de silhouette moyen le plus élevé parmi les différentes méthodes utilisées, c’est elle qui a été retenue pour la

caractérisation des profils des patients entrés en traitement en 2021 ainsi que pour la comparaison de ces profils à ceux observés sur la période 2015 à 2020.

Notre étude a ainsi permis d'identifier deux principaux clusters parmi les patients. Le premier cluster, représentant la majorité des patients, se compose de personnes ayant une dépendance relativement moins sévère aux substances. Ceux-ci sont principalement envoyés en traitement par la justice et déclarent vivre en couple, résider dans des domiciles fixes et avoir un niveau d'instruction secondaire. En revanche, le deuxième cluster se compose de patients ayant une dépendance plus sévère aux substances. Ils sont principalement référés par des membres de leur famille ou des hôpitaux, et leurs principales sources de revenus sont les allocations familiales. De plus, ils ont indiqué avoir vécu avec des amis au cours du mois précédent.

Nos analyses ont également montré des différences significatives entre ces 2 clusters. Ces différences portent sur des facteurs tels que l'âge lors du premier usage du cannabis, le niveau d'éducation, la situation de logement et la dépendance à d'autres substances.

Bien que les moyennes d'âge des patients des deux clusters soient similaires, il existe une différence significative dans l'âge lors du premier usage du cannabis. Les patients du cluster 2 ont commencé à consommer du cannabis à un âge plus précoce que ceux du cluster 1. En ce qui concerne la dépendance aux substances, le cluster 1 présente une faible fréquence de problèmes liés à la consommation de substances autres que le cannabis. En revanche, le cluster 2 se distingue par une plus grande prévalence de problèmes de dépendance à la cocaïne en poudre, aux amphétamines, à la MDMA, à la méphédrone, aux benzodiazépines, à la kétamine et à l'alcool.

De plus, l'analyse temporelle des clusters sur la période allant de 2015 à 2020 a révélé la persistance de deux clusters principaux, confirmant ainsi les résultats obtenus pour l'année 2021. Une légère variation a été observée sur l'année 2019 où trois clusters distincts ont été identifiés, les deux premiers étant similaires à ceux observés en 2021. Le troisième cluster, représentant une minorité de patients, se compose de patients pré-pensionnés ou pensionnés, manifestant également une propension élevée à recourir à l'injection de substances. Ces résultats suggèrent ainsi une stabilité des profils des patients dépendants du cannabis et entrants en traitement pour la première fois ; aucune évolution significative n'a été observée sur la période allant 2015 à 2021.

Cette étude apporte une lumière sur les profils des personnes demandant un traitement pour leur dépendance au cannabis en Belgique. Les deux clusters identifiés ont des implications significatives dans un contexte plus large sur le plan national ainsi que pour la santé publique en Belgique.

Le fait que le premier cluster représente la majorité des patients, avec une dépendance moins sévère aux drogues illégales et à l'alcool, suggère qu'il existe une prévalence élevée de personnes présentant des problèmes de consommation de substances moins graves. Cela pourrait indiquer que la dépendance au cannabis en Belgique est souvent associée à une utilisation moins problématique d'autres drogues et d'alcool. De plus, le fait que l'origine du début de traitement de ces patients soit lié à l'intervention de la justice souligne le lien entre la justice et la demande de traitement pour la dépendance au cannabis en Belgique. Cela indique que les programmes de traitement judiciaire sont utilisés comme moyen de

faire face à la dépendance au cannabis dans certains cas. Cette observation rejoint les résultats obtenus dans des études précédentes sur les personnes entrant en traitement pour une dépendance au cannabis [Antoine *et al.*, 2022, Gisle et Drieskens, 2019]. La coordination entre le système judiciaire et les autorités sanitaires serait donc nécessaire, afin de garantir une approche intégrée et efficace.

D'autre part, le second cluster identifié, composé de patients présentant une dépendance polyproblématique aux substances et ayant souvent recours à l'injection de substances souligne la présence d'une population plus vulnérable et à haut risque, nécessitant des interventions spécifiques et des approches de traitement adaptées. Ces patients peuvent avoir besoin d'un soutien plus intensif, y compris des programmes de réduction des méfaits et des services de santé mentale complémentaires. De plus, le fait que ces patients aient débuté le traitement sous incitation de familiale et des hôpitaux met en lumière l'importance des facteurs sociaux et familiaux dans le processus de demande de traitement pour la dépendance au cannabis et souligne nécessité un renforcer les programmes de prévention et d'intervention précoce, ainsi que des ressources pour aider les familles à reconnaître les signes précoces de la dépendance et à accéder à un traitement adéquat.

En ce qui concerne l'évolution stable des clusters dans le temps, cela indique que les profils de dépendance et les besoins de traitement des patients restent relativement constants.

La présente étude est sujette à certaines limitations et plusieurs améliorations peuvent y être apportées. Mentionnons en premier lieu les techniques mises en oeuvre pour identifier les clusters. Celles ci sont basées en effet sur les centroïdes et le clustering hiérarchique. Bien que ces techniques soient populaires, elles ne fournissent pas, étant données les caractéristiques d'un patient, la probabilité qu'il appartienne à un cluster particulier. Il pourrait donc être intéressant d'explorer les résultats d'un algorithme supplémentaire qui peut fournir des probabilités d'appartenance à un cluster. Un algorithme intéressant qui pourrait être utilisé est le modèle de mélange gaussien [Biernacki, 2009]. Plutôt que de catégoriser les patients de manière rigide dans un seul cluster, cette approche permettra de nuancer les résultats en présentant les probabilités d'appartenance associées. Cela permet une description plus nuancée des profils des patients et peut orienter les analyses supplémentaires ou les décisions relatives aux traitements et aux interventions par le personnel en charge des patients. Ensuite, un autre axe d'amélioration de ce travail consisterait à ajouter une mesure supplémentaire de l'évaluation de la qualité des clusters, en complément du score de silhouette. En effet, bien que le score de silhouette soit largement utilisé pour évaluer la cohésion et la séparation des clusters, des méthodes telles que l'indice de Calinski-Harabasz ou l'indice de Dunn pourraient être utilisées [Wang et Xu, 2019, Lallich et Lenca, 2015] pour apporter des informations complémentaires sur la qualité des clusters, telles que la compacité et la séparation des clusters. Enfin, une amélioration essentielle à apporter à cette étude est de poursuivre le suivi des profils observés sur la période post-COVID afin de compenser les limitations rencontrées lors de la collecte des données pendant cette période. En effet, la situation qui a prévalu en 2020 a fortement impacté l'offre de service et la demande de traitement des consommateurs de drogues ainsi que la qualité des données de l'indicateur TDI. Comme le souligne Sciensano dans son rapport sur l'enregistrement TDI en Belgique pour l'année 2020, "le rapportage correct et complet des données a pu

poser un problème pour les professionnels et les données reçues ne reflètent donc pas la demande de traitement réelle" [Antoine, 2021]. Ce facteur peut avoir eu un impact sur la représentativité des profils identifiés sur la période 2020-2021 et la compréhension de l'évolution des clusters dans le temps. Une amélioration de la qualité des données disponibles permettra de prendre en compte les éventuelles modifications dans les clusters afin de valider les résultats obtenus dans cette étude sur la période post-covid, de mieux comprendre les impacts de la pandémie sur la dépendance au cannabis et mieux évaluer les besoins de traitement.

Dans une perspective de recherche future, il serait intéressant de poursuivre l'exploration des clusters identifiés en se concentrant sur l'identification de sous-groupes au sein de ces clusters. Cette approche permettrait de mieux comprendre les caractéristiques et les besoins spécifiques des patients au sein de chaque cluster, et ainsi d'adapter les interventions et les traitements de manière plus ciblée.

Bibliographie

- [Ahmad et Khan, 2019] AHMAD, A. et KHAN, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *Ieee Access*, 7:31883–31902.
- [Antoine, 2021] ANTOINE, J. (2021). L’enregistrement tdi en belgique : Rapport annuel, année d’enregistrement 2020. *Bruxelles, Belgique : Sciensano. Numéro de dépôt légal : D/2021/14.440/77*.
- [Antoine, 2022] ANTOINE, J. (2022). L’enregistrement tdi en belgique : Rapport annuel, année d’enregistrement 2021. *Bruxelles, Belgique : Sciensano. 105p. Numéro de dépôt légal : D/2022.14.440/65*.
- [Antoine et al., 2016] ANTOINE, J., DE RIDDER, K., PLETTINCKX, E., BLANCKAERT, P. et GREMEAUX, L. (2016). Treatment for substance use disorders : the belgian treatment demand indicator registration protocol. *Archives of Public Health*, 74(1):1–8.
- [Antoine et al., 2022] ANTOINE, J., HOGGE, M., DE DONDER, E., VERSTUYF, G., PLETTINCKX, E. et GREMEAUX, L. (2022). Which drugs of choice are on the rise or in decline? a trend analysis of belgium’s treatment settings (2015-2019). *Drugs, Habits and Social Policy*, (ahead-of-print).
- [Biernacki, 2009] BIERNACKI, C. (2009). Pourquoi les modèles de mélange pour la classification? *Monde des Util. Anal. Données*, 40:1–22.
- [Cadario et Parguel, 2014] CADARIO, R. et PARGUEL, B. (2014). Reconsidérer la discrétisation des variables quantitatives : Vers une nouvelle analyse de modération en recherche expérimentale. *Recherche et Applications en Marketing (French Edition)*, 29(4):120–140.
- [Coquillard, 2018] COQUILLARD, P. (2018). «méthodes de classification» (université de nice, france).
- [Dinh et al., 2019] DINH, D.-T., FUJINAMI, T. et HUYNH, V.-N. (2019). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In *Knowledge and Systems Sciences : 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20*, pages 1–17. Springer.
- [Dubey et Robin, 2006] DUBY, C. et ROBIN, S. (2006). Analyse en composantes principales. *Institut National Agronomique, Paris-Grignon*, 80:53.
- [Fu et al., 2004] FU, L., GOH, D. H.-L. et FOO, S. S.-B. (2004). The effect of similarity measures on the quality of query clusters. *Journal of information Science*, 30(5):396–407.
- [Gisle et Drieskens, 2019] GISLE, L. et DRIESKENS, S. (2019). Enquête de santé 2018 : Usage des drogues.

- [Huang, 1998] HUANG, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3):283–304.
- [Humaira et Rasyidah, 2020] HUMAIRA, H. et RASYIDAH, R. (2020). Determining the appropriate cluster number using elbow method for k-means algorithm. In *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*.
- [Lallich et Lenca, 2015] LALLICH, S. et LENCA, P. (2015). Indices de qualité en clustering. In *Journée thématique : clustering et co-clustering*. Société française de classification.
- [Lebart et al., 1995] LEBART, L., MORINEAU, A. et PIRON, M. (1995). *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris.
- [MacQueen, 1967] MACQUEEN, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA.
- [Malik et Tuckfield, 2019] MALIK, A. et TUCKFIELD, B. (2019). *Applied unsupervised learning with R : Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA*. Packt Publishing Ltd.
- [Mamat et al., 2018] MAMAT, A. R., MOHAMED, F. S., MOHAMED, M. A., RAWI, N. M. et AWANG, M. I. (2018). Silhouette index for determining optimal k-means clustering on images in different color models. *Int. J. Eng. Technol*, 7(2):105–109.
- [Pagès, 2002] PAGÈS, J. (2002). Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de statistique appliquée*, 50(4):5–37.
- [Palani et al., 2021] PALANI, K., STYNES, P. et PATHAK, P. (2021). Clustering techniques to identify low-engagement student levels. In *CSEDU (2)*, pages 248–257.
- [Palm, 1998] PALM, R. (1998). L’analyse en composantes principales : principes et applications. *Notes de Statistique et d’Informatique*, (2).
- [Palm, 2007] PALM, R. (2007). L’analyse des correspondances multiples : principes et application. *Notes de Statistique et d’Informatique*, (2).
- [Rousseeuw, 1987] ROUSSEEUW, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- [Schönnesson et al., 2008] SCHÖNNESSON, L. N., ATKINSON, J., WILLIAMS, M. L., BOWEN, A., ROSS, M. W. et TIMPSON, S. C. (2008). A cluster analysis of drug use and sexual hiv risks and their correlates in a sample of african-american crack cocaine smokers with hiv infection. *Drug and alcohol dependence*, 97(1-2):44–53.
- [Wang et Xu, 2019] WANG, X. et XU, Y. (2019). An improved index for clustering validation based on silhouette index and calinski-harabasz index. In *IOP Conference Series : Materials Science and Engineering*, volume 569, page 052024. IOP Publishing.
- [Yong et al., 2013] YONG, A. G., PEARCE, S. et al. (2013). A beginner’s guide to factor analysis : Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2):79–94.
- [Zhou et Gao, 2014] ZHOU, H. B. et GAO, J. T. (2014). Automatic method for determining cluster number based on silhouette coefficient. In *Advanced materials research*, volume 951, pages 227–230. Trans Tech Publ.

Annexe A

Analyses complémentaires

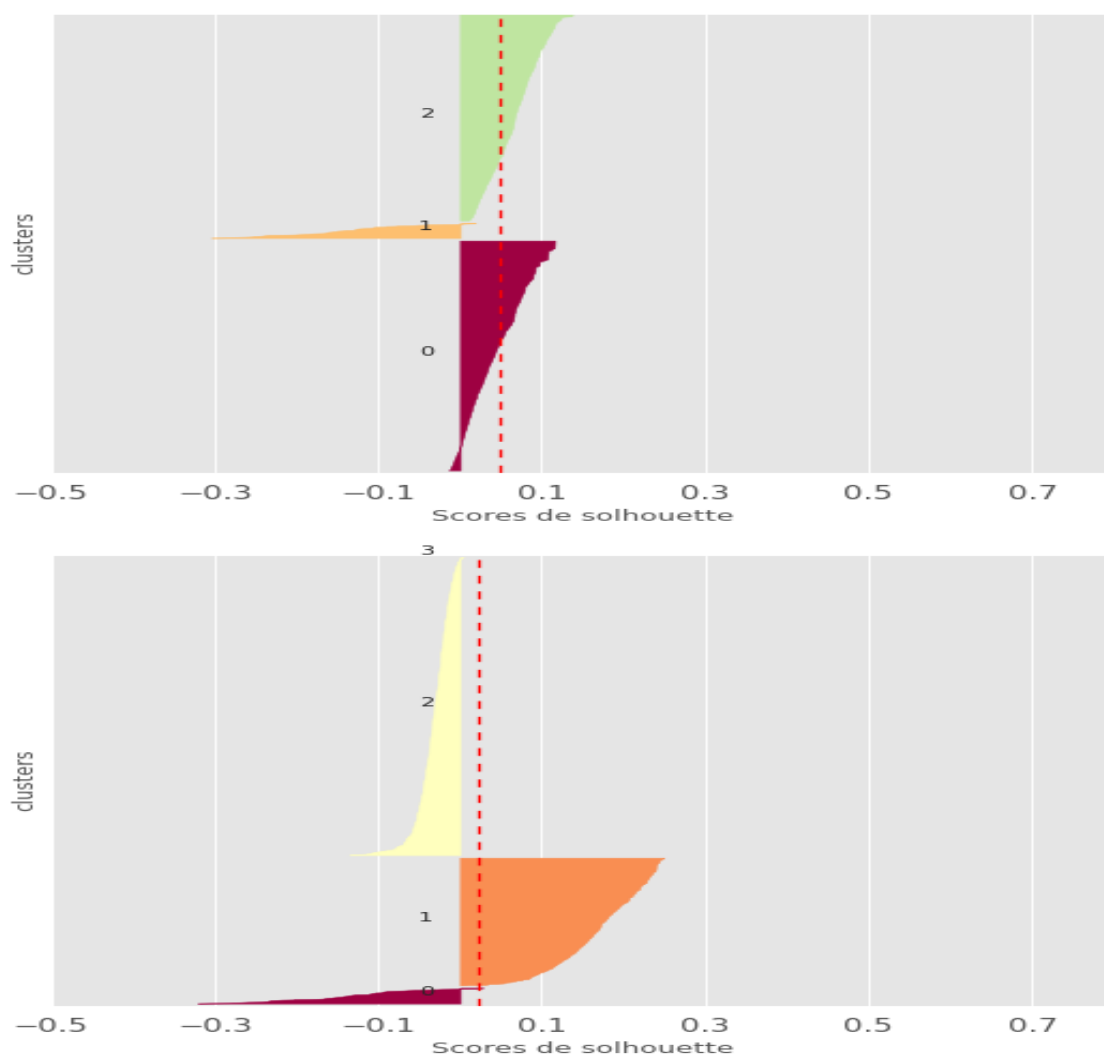


FIGURE A.1 – Analyse des tracés de silhouette(pour $k \in \{3, 4, 5, 6, 7\}$)

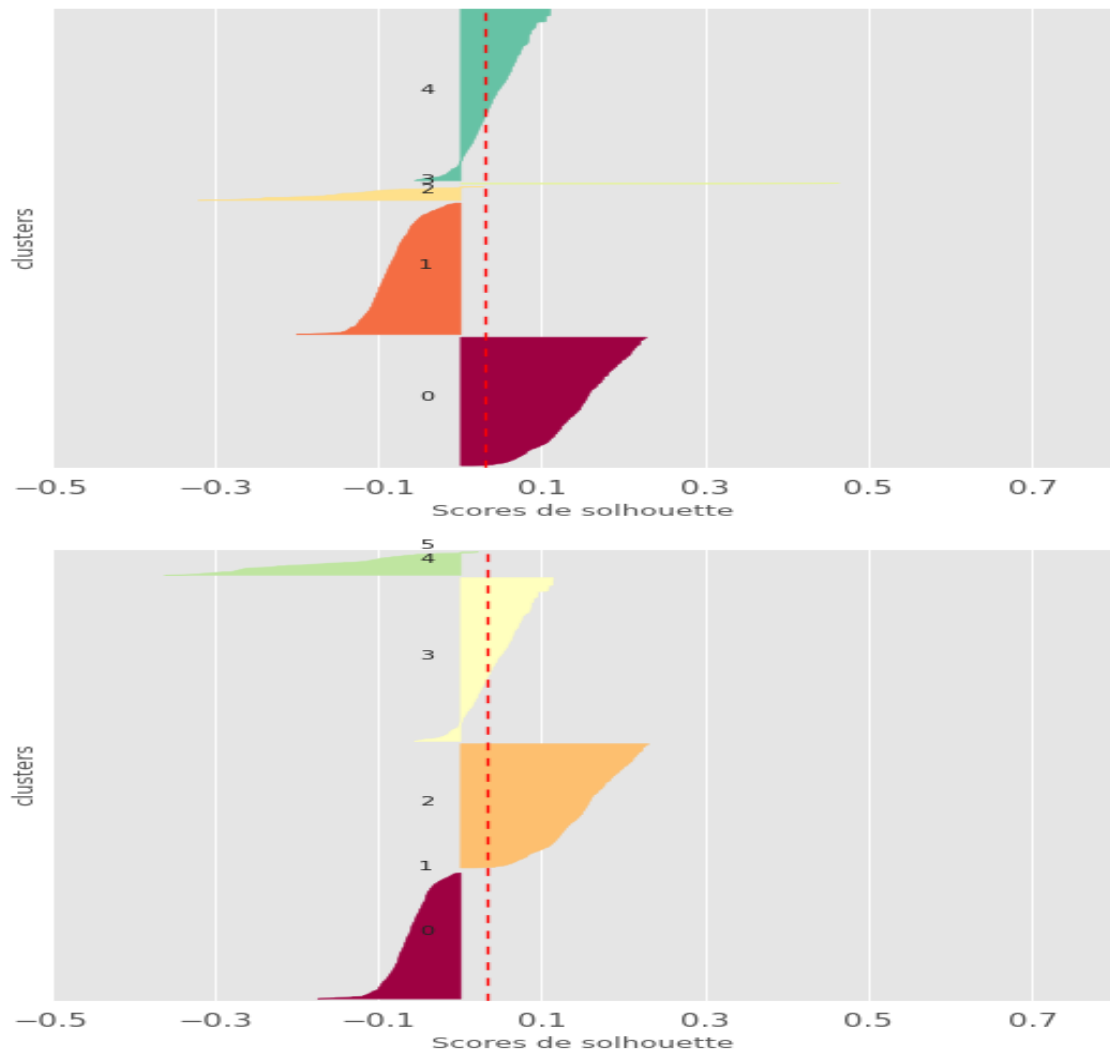


FIGURE A.2 – Analyse des tracés de silhouette(pour $k \in \{3, 4, 5, 6, 7\}$)

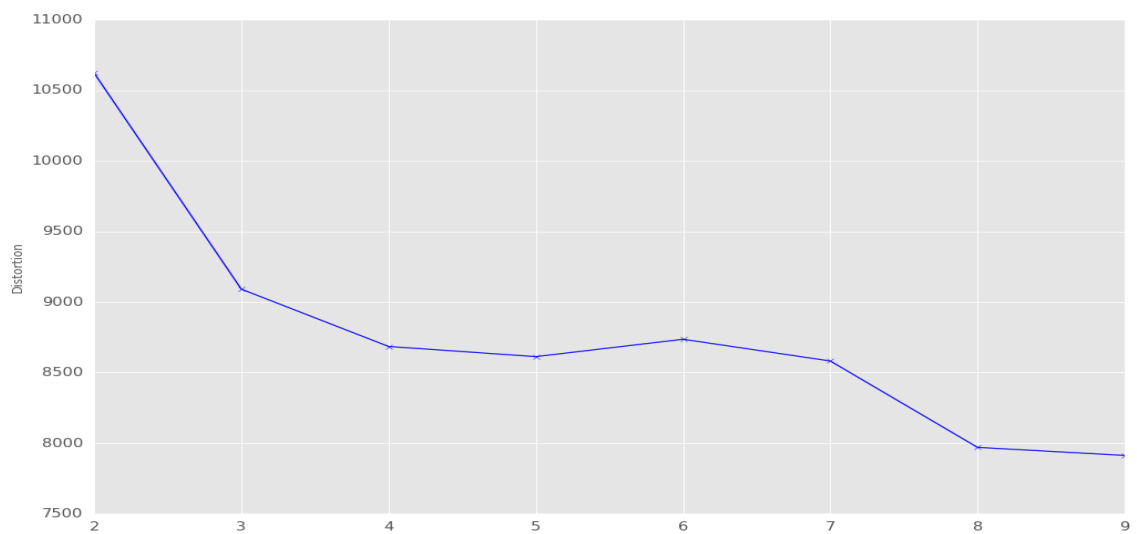


FIGURE A.3 – Elbow plot : kmodes

TABLE A.1 – Répartition des patients en fonction des caractéristiques sociodémographiques

Variables		effectif	fréquence(%)
CD_SEX			
	Femme	368	19.92
	Homme	1479	80.08
CD_LVN_STAT_WITH_WHOM			
	Autre	16	0.87
	Avec amis	78	4.22
	Avec Famille	95	5.14
	Avec parent(s)	859	46.51
	En couple	319	17.27
	Seul	480	25.99
CD_LVN_CHILDREN			
	Non	1497	81.05
	Oui	350	18.95
CD_LVN_STAT_WHERE			
	Autre	10	0.54
	Domicile fixe	1654	89.55
	Institution	73	3.95
	Logements variables	69	3.74
	Prison	25	1.35
	Rue	16	0.87
CD_EDUCATION_LEVEL			
	aucun	34	1.84
	Autre	17	0.92
	primaire	518	28.05
	secondaire	1167	63.18
	supérieure	111	6.01
CD_LABOUR_STAT			
	Au foyer	21	1.14
	Autre	22	1.19
	Aux études	671	36.33
	Chômage	334	18.08
CD_LABOUR_STAT			
	Incapacité travail	223	12.07
	Occasionnel	48	2.60
	Pensionné/pré-pensionné	2	0.11
	Régulier	526	28.48
	Alloc.Chômage	180	9.75
CD_INCOME_STAT			
	Alloc.familiales	44	2.38
	Aucun revenu	674	36.49
	Autre	11	0.60
	Bourse d'études	2	0.11
	CPAS70	140	7.58
	Indemnités	238	12.89
	Pension	4	0.22
	Salaire	554	29.99

TABLE A.2 – Répartition des patients en fonction du profil d'addiction

variables		effectif	fréquence(%)
CD_FREQ_USE			
	2-3j/sem	192	10.40
	4-6j/sem	143	7.74
	Max 1j/sem	134	7.26
	Pas le mois précédent	287	15.54
	Tous les jours	1091	59.07
CD_INJECTING_STAT			
	Non	1828	98.97
	Oui	19	1.03
CD_REFERRAL			
	Ami	43	2.33
	Autre	70	3.79
	Autre service médical	167	9.04
	Centre pour toxicomanes	19	1.03
	Famille	177	9.58
	Hôpital	80	4.33
	Individuel	451	24.42
	Justice	753	40.77
	Médecin	87	4.71

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
Score de silhouette moyen	0.548	0.489	0.489	0.492	0.026	0.029

TABLE A.3 – Scores de silhouette moyen en fonction du nombre de clusters : année 2015

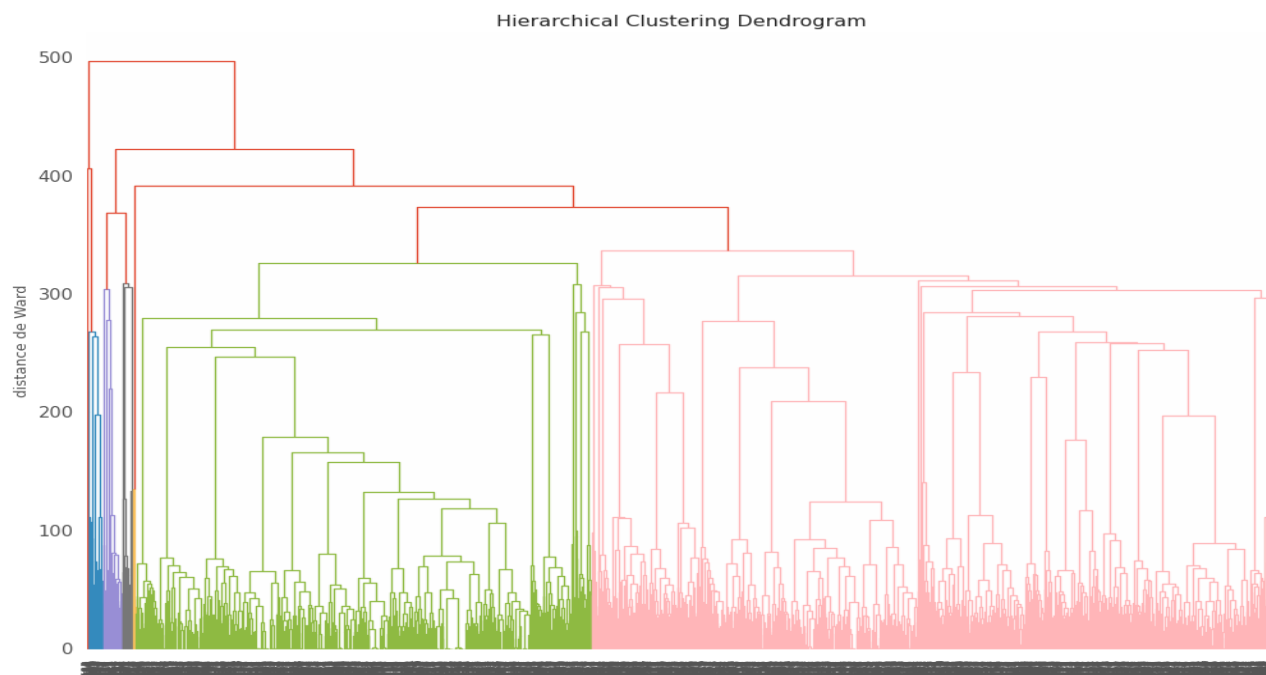


FIGURE A.4 – Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2015

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
Score de silhouette moyen	0.532	0.512	0.471	0.032	0.036	0.035

TABLE A.4 – Score de silhouette moyen en fonction du nombre de clusters : année 2016

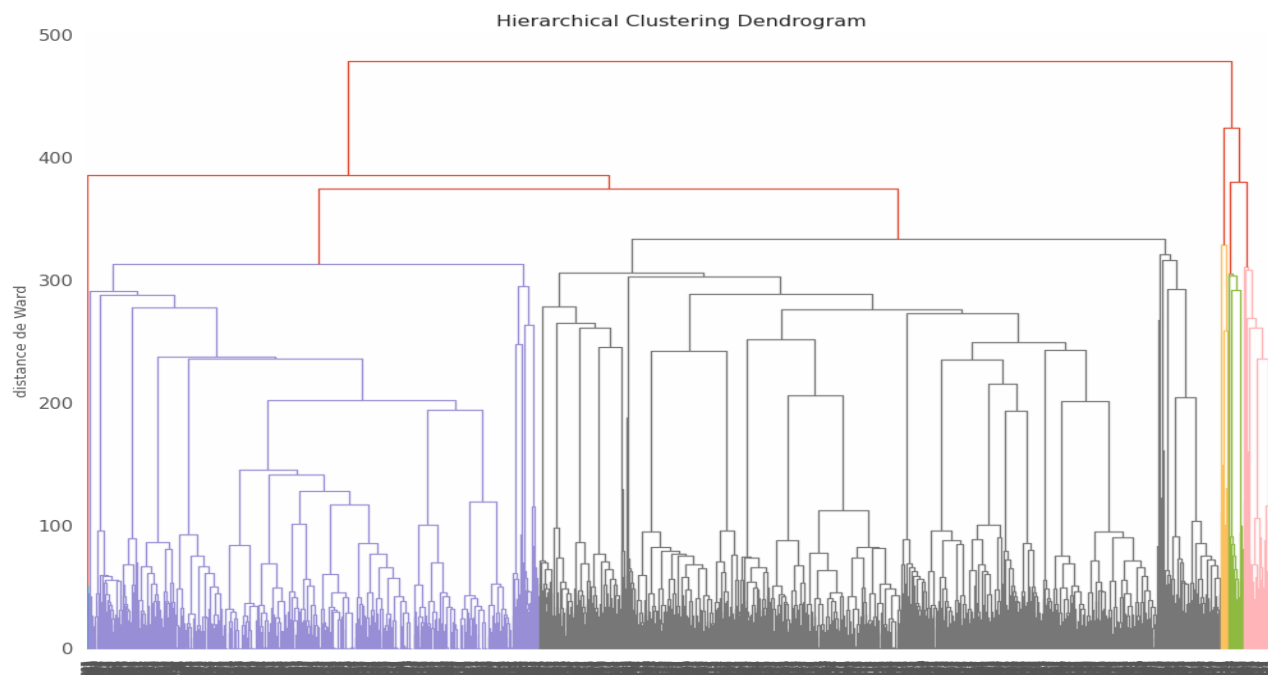


FIGURE A.6 – Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour la première fois en 2017

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
Score de silhouette moyen	0.462	0.458	0.458	0.400	0.057	0.060

TABLE A.6 – Score de silhouette moyen en fonction du nombre de clusters : année 2018

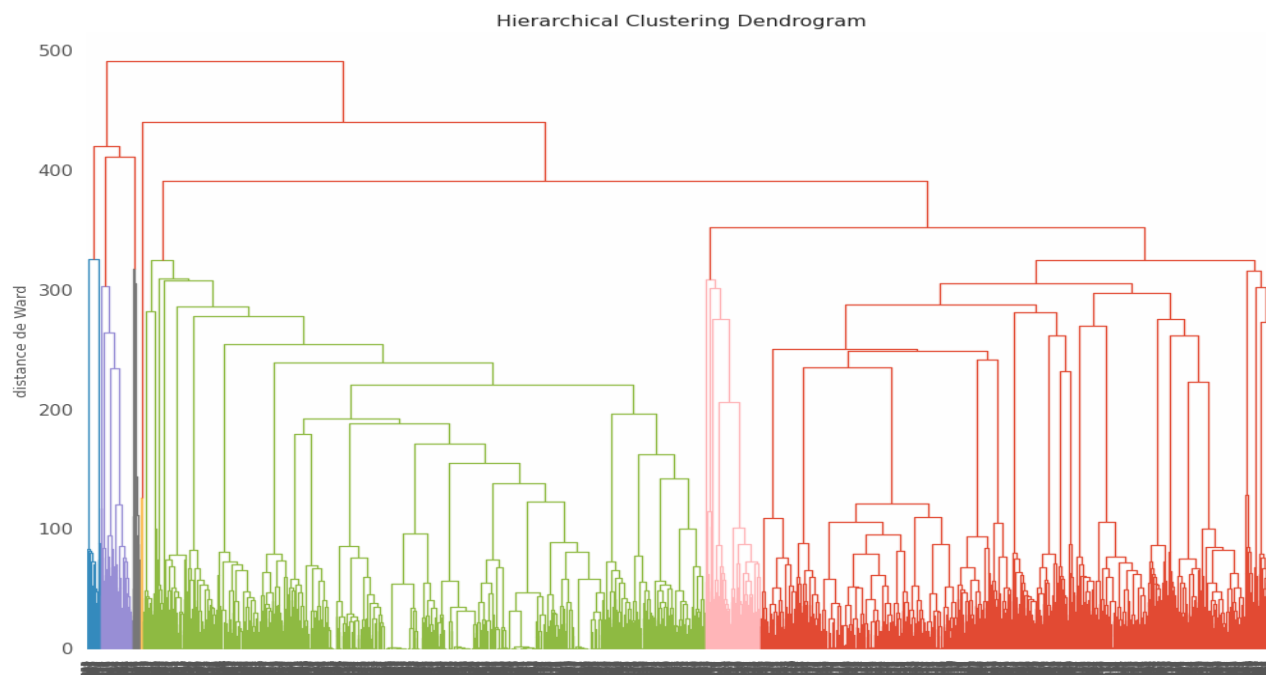


FIGURE A.7 – Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2018

	Nombre de clusters					
	k=2	k=3	k=4	k=5	k=6	k=7
Score de silhouette moyen	0.486	0.422	0.423	0.426	0.271	0.249

TABLE A.7 – Score de silhouette moyen en fonction du nombre de clusters : année 2020

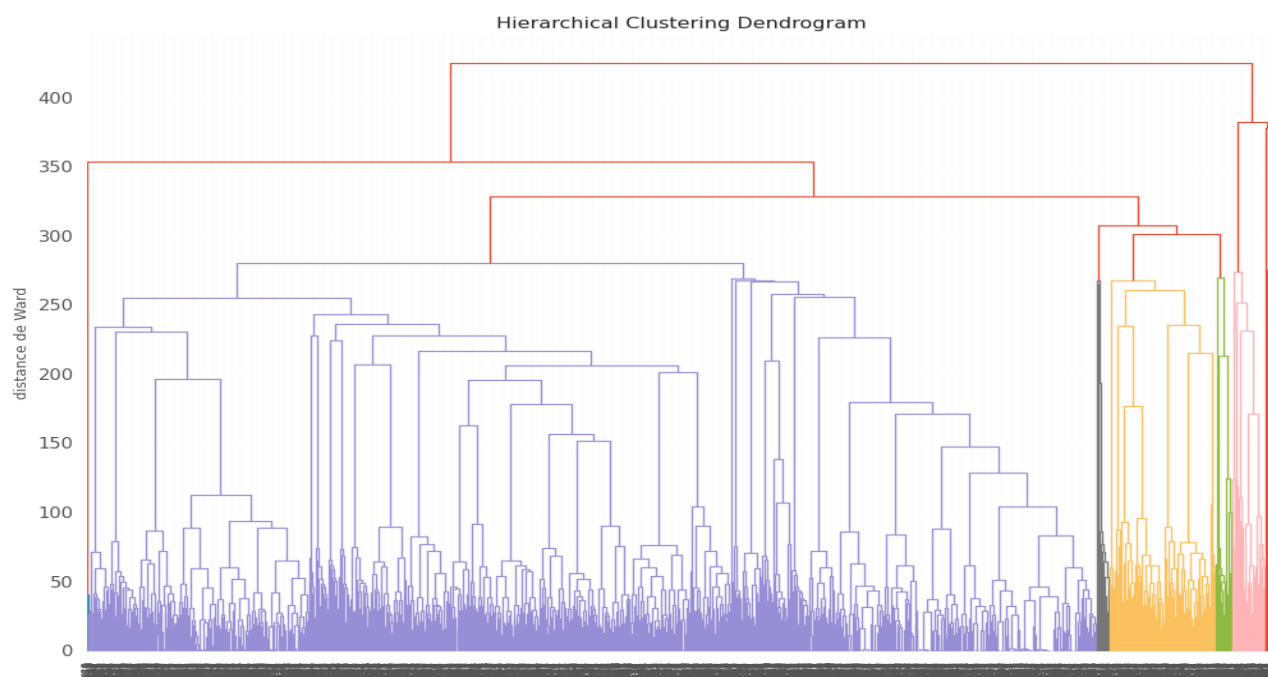


FIGURE A.8 – Dendrogramme de la Classification Ascendante Hiérarchique des patients entrés en traitement pour le première fois en 2020

Annexe B

Code python

```
1 # author: Keline Laure Meniodem
2
3 ## Installation of libraries
4 from IPython.display import HTML
5 import base64, io
6
7 !pip install --upgrade rpy2
8 ! pip install matplotlib==3.7.1
9 import matplotlib
10 print('Matplotlib version: ', matplotlib.__version__) # >= 2.0.0
11 # Commented out IPython magic to ensure Python compatibility.
12 # %%capture
13 # ! pip install seaborn
14
15 ! pip install gower
16 ! pip install prince
17 ! pip install kmodes
18 ! pip install yellowbrick
19
20 import sklearn
21 #print(sklearn.__version__ > 1.1.2)
22
23 !pip install researchpy
24 #!pip uninstall stargazer -y
25 !pip install git+https://github.com/mwburke/stargazer.git
26 pip install scientisttools
27
28 ## Importing the libraries
29
30 import pandas as pd
31 import numpy as np
32 import matplotlib.pyplot as plt
33 import gower
34 import prince
35 import warnings
```

```

36 import rpy2
37 from rpy2 import robjects
38 from stargazer.stargazer import Stargazer
39 from sklearn.cluster import KMeans
40 from sklearn.datasets import make_blobs
41 from sklearn.impute import SimpleImputer
42 from yellowbrick.cluster import KElbowVisualizer,
    SilhouetteVisualizer
43 from scipy.cluster.hierarchy import dendrogram, linkage
44 from sklearn.metrics import silhouette_samples, silhouette_score
45 from kmodes import kprototypes
46 from tabulate import tabulate
47 import subprocess
48 import matplotlib.cm as cm
49 import plotly.express as px
50 import scipy.stats as ss
51 # %matplotlib inline
52 import statsmodels.formula.api as smf
53 from statsmodels.stats.multicomp import pairwise_tukeyhsd
54 from statsmodels.stats.multicomp import MultiComparison
55 import scipy.stats as stats
56 import researchpy as rp
57 from scipy.stats import chi2_contingency
58 import statsmodels.api as sm
59 from stargazer.stargazer import Stargazer
60 from sklearn.cluster import KMeans
61 from sklearn.decomposition import PCA
62 from sklearn.preprocessing import StandardScaler, OneHotEncoder
63 from kmodes.kprototypes import KPrototypes
64 from sklearn import preprocessing
65 from kmodes.kmodes import KModes
66 import scipy.cluster.hierarchy as shc
67 from sklearn.cluster import AgglomerativeClustering
68 from altair import hconcat
69 from sklearn.preprocessing import StandardScaler, MinMaxScaler
70 from sklearn.linear_model import LogisticRegression
71 from sklearn.metrics import silhouette_score
72 from IPython.display import display, HTML
73 from tqdm import tqdm
74 import scipy.stats as stats
75 from termcolor import colored
76
77 warnings.filterwarnings('ignore')
78 plt.style.use('fivethirtyeight')
79
80 import seaborn as sns
81 pd.set_option("display.max_columns", None)
82
83

```

```

84 ## Importing the data
85
86 df = pd.read_sas('/content/drive/MyDrive/tdidataset.sas7bdat')
87 df.head(2)
88
89 ## Data preprocessing
90
91 df.duplicated().sum() #Duplicate data
92 df.drop_duplicates(inplace = True)
93 df.duplicated().sum() # Allows to ensure that there are no more
    duplicate values
94 df['CD_LVN_CHILDREN'].unique()
95 df['CD_LVN_STAT_WITH_WHOM'].unique()
96 # replace all unwanted values with NaN
97 df = df.replace({88: 'Autre', 99: np.NaN, -1:np.NaN, -2:np.NaN,
    9999:np.NaN})
98 df['CD_LVN_CHILDREN'] = df['CD_LVN_CHILDREN'].replace({0:np.NaN})
99 df['CD_LVN_STAT_WITH_WHOM'] = df['CD_LVN_STAT_WITH_WHOM'].replace
    ({0:np.NaN})
100 df['CD_ROUTE_ADMIN'] = df['CD_ROUTE_ADMIN'].replace({0:np.NaN})
101 df['CD_FREQ_USE'] = df['CD_FREQ_USE'].replace({0:np.NaN})
102 # Inspection of variables through descriptive statistics
103
104 df.describe()
105 # Calculating the number of null values in the "CD_MAIN_SUBST"
    column.
106 df['CD_MAIN_SUBST'].isnull().sum()
107
108 ### Extraction of cannabis dependents
109 # Recoding the variable "CD_MAIN_SUBST"
110 def my_recode(main_substance):
111     if main_substance in range(10,16):
112         return "opiacés"
113     elif main_substance in range(20,24):
114         return "cocaïne"
115     elif main_substance in range(30,36):
116         return "stimulants"
117     elif main_substance in range(40,45):
118         return "hypnotiques"
119     elif main_substance in range(70,74):
120         return "cannabis"
121     elif main_substance == 80:
122         return "alcool"
123     elif main_substance in range(50,54):
124         return "hallucinogènes"
125     else:
126         return "autres"
127 # Using the function "my_recode" to modify the variable "
    main_subst"

```

```

128 df['CD_MAIN_SUBST_recode'] = df['CD_MAIN_SUBST'].apply(my_recode)
129 # Extracting patients for whom cannabis is the main substance
130 df_cannabis = df.query('CD_MAIN_SUBST_recode=="cannabis"')
131 df_cannabis.head(2)
132 ### Sélection des variables d'intérêt
133 # liste des variables sélectionnées
134 list_selected_var = ['CD_SEX', 'NR_AGE',
135                     'CD_LVN_STAT_WHERE', 'CD_LVN_STAT_WITH_WHOM', '
136                     CD_LVN_CHILDREN',
137                     'CD_EDUCATION_LEVEL', 'CD_LABOUR_STAT', 'CD_INCOME_STAT',
138                     'CD_REFERRAL', 'CD_PREV_TREAT', 'FL_OPIATES',
139                     'FL_HEROIN', 'FL_BUPRENORPHINE', 'FL_METHADONE', '
140                     FL_FENTANYL',
141                     'FL_OTHER_OPIATES', 'FL_COCAINE', 'FL_POWDER_COCAINE', '
142                     FL_CRACK',
143                     'FL_OTHER_COCAINE', 'FL_STIMULANTS', 'FL_AMPHETAMINES',
144                     'FL_METHAMPHETAMINES', 'FL_MDMA', 'FL_MEPHEDRONE',
145                     'FL_OTHER_STIMULANTS', 'FL_HYPNOTICS', 'FL_BARBITURATES',
146                     'FL_BENZODIAZEPINES', 'FL_GHB', 'FL_OTHER_HYPNOTICS',
147                     'FL_HALLUCINOGENS', 'FL_LSD', 'FL_KETAMINE', '
148                     FL_OTHER_HALLUCINOGENS',
149                     'FL_VOLATILE', 'FL_CANNABIS', 'FL_MARIJUANA', 'FL_HASH',
150                     'FL_OTHER_CANNABIS', 'FL_ALCOHOL', 'FL_OTHER_SUBSTANCE',
151                     'MS_AGE_FIRST_USE', 'CD_FREQ_USE',
152                     'CD_ROUTE_ADMIN',
153                     'NR_DATE_START_TREAT', 'CD_INJECTING_STAT'] # Adding the
154                             injection status
155
156 # Data containing only the selected variables
157 df_cannabis_select = df_cannabis[list_selected_var]
158 df_cannabis_select.head(2)
159 len(list_selected_var)
160
161 ### Inspection of missing values
162 df_cannabis_interest = df_cannabis_select.copy()
163 df_cannabis_interest.head(2)
164 # function that displays the percentage of missing values in each
165     column of the dataframe
166 def missing(dff):
167     print (round((dff.isnull().sum() * 100/ len(dff)),2).
168             sort_values(ascending=False))
169
170 print(missing(df_cannabis_interest))
171
172 # Function to remove columns with a specified threshold of
173     missing values
174 def rmissingvaluecol(dff, threshold):
175     l = []

```

```

168     l = list(dff.drop(dff.loc[:,list(((100*(dff.isnull().sum()/len
      (dff.index)) >= threshold))].columns, 1).columns.values)
169     print("# Columns having more than %s percent missing values:
      "%threshold, (dff.shape[1] - len(l)))
170     print("Columns:\n", list(set(list((dff.columns.values))) -
      set(l)))
171     return l
172
173 # Using the function rmissingvaluecol to remove columns with more
      than 80% missing values
174 l=rmissingvaluecol(df_cannabis_interest, 80)
175
176 # Database containing the remaining variables
177 df_cannabis_interest = df_cannabis_interest[l]
178 df_cannabis_interest.duplicated().sum()
179 df_cannabis_interest.drop_duplicates(inplace=True)
180 df_cannabis_interest.duplicated().sum()
181
182
183 # Changing the data types of the variables
184
185 df_cannabis_interest['NR_DATE_START_TREAT'] = pd.to_datetime(
      df_cannabis_interest['NR_DATE_START_TREAT'])
186 cols = ['NR_AGE', 'NR_DATE_START_TREAT', 'MS_AGE_FIRST_USE']
187 for col in list(df_cannabis_interest.columns):
188     if col not in cols:
189         #df_cannabis_interest[col] = df_cannabis_interest[col].astype
          ('int')
190         df_cannabis_interest[col] = df_cannabis_interest[col].astype(
          'category')
191
192 # Imputation of missing values.
193
194 imputer_num = SimpleImputer(strategy='median')
195 imputer_cat= SimpleImputer(strategy='most_frequent')
196 # Verifying the shape of the column "CD_SEX"
197 print(df_cannabis_interest['CD_SEX'].shape)
198 df_cannabis_interest['CD_SEX'] = df_cannabis_interest['CD_SEX'].
      transform(lambda x: x.fillna(x.mode()[0]))
199 cols=['NR_AGE', 'MS_AGE_FIRST_USE', 'NR_DATE_START_TREAT']
200 for col in list(df_cannabis_interest.columns):
201     if col not in cols:
202         df_cannabis_interest[col] = df_cannabis_interest.groupby(['
          CD_SEX'])[col].transform(lambda x: x.fillna(x.mode()[0]))
203
204 # Imputing missing values for numeric variables (age and age at
      treatment initiation) by sexe
205 df_cannabis_interest[['NR_AGE',

```

```

206         'MS_AGE_FIRST_USE']] = df_cannabis_interest
207         .groupby(['CD_SEX'])[['NR_AGE',
208         'MS_AGE_FIRST_USE']].transform(lambda x: x.
209         fillna(x.median()))
210 df_cannabis_interest.duplicated().sum()
211 df_cannabis_interest.drop_duplicates(inplace=True)
212 df_cannabis_interest.duplicated().sum()
213 print(missing(df_cannabis_interest))
214
215 ## Recoding the variables
216
217 dict_sex = {1: 'Homme', 2: 'Femme'}
218 dict_reg = {1: 'Flandre', 2: 'Wallonnie', 3: 'Bruxelles'}
219 dict_educLevel = {1: 'aucun', 2: 'primaire', 3: 'secondaire', 4: '
    supérieure'}
220 dict_lvn_where = {1: 'Domicile fixe', 2: 'Logements variables', 3:
    'Rue', 4: 'Institution', 5: 'Prison', 6: 'Autre'}
221 dict_lvn_whith_whom = {1: 'Seul', 2: 'En couple', 3: 'Avec parent(s)
    ', 4: 'Avec Famille', 5: 'Avec amis', 6: 'Autre'}
222 dict_lvn_children = {1: 'Oui', 2: 'Non'}
223 dict_labour_stat = {1: 'Régulier', 2: 'Occasionnel', 3: 'Chômage'
    , 4: 'Aux études', 5: 'Incapacité travail',
224     6: 'Au foyer', 7: 'Pensionné/pré-pensionné', 8:
    'Autre'}
225 dict_income_stat = {1: 'Salaire', 2: 'Alloc.Chômage', 3: 'Pension'
    , 4: 'Indemnités', 5: 'CPAS', 6: 'Alloc.familiales',
226     7: 'Bourse d\'études', 8: 'Aucun revenu', 9: '
    Autre'}
227
228 dict_referral = {1: 'Individuel', 2: 'Famille', 3: 'Ami', 4: 'Médecin'
    , 5: 'Centre toxicomanes', 6: 'Hôpital',
229     7: 'Autre service médical', 8: 'Justice', 9: 'Autre'
    }
230 dict_route_admin = {1: 'Injection', 2: 'Fumer/inhaler', 3: 'Manger/
    boire', 4: 'Snifer', 5: 'Autre'}
231 dict_freq_use = {1: 'Pas le mois précédent', 2: 'Max 1j/sem', 3: '
    2-3j/sem', 4: '4-6j/sem', 5: 'Tous les jours'}
232 dict_prev_treat = {1: 'Oui', 2: 'Non'}
233
234 # Types of substances
235 dict_FL_OPIATES = {1: 'Oui', 0: 'Non'}
236 dict_FL HEROIN = {1: 'Oui', 0: 'Non'}
237 dict_FL BUPRENORPHINE = {1: 'Oui', 0: 'Non'}
238 dict_FL METHADONE = {1: 'Oui', 0: 'Non'}
239 dict_FL FENTANYL = {1: 'Oui', 0: 'Non'}
240 dict_FL OTHER_OPIATES = {1: 'Oui', 0: 'Non'}
241 dict_FL COCAINE = {1: 'Oui', 0: 'Non'}
242 dict_FL POWDER_COCAINE = {1: 'Oui', 0: 'Non'}
243 dict_FL CRACK = {1: 'Oui', 0: 'Non'}

```

```

242 dict_FL_OTHER_COCAINE = {1: 'Oui', 0: 'Non'}
243 dict_FL_STIMULANTS = {1: 'Oui', 0: 'Non'}
244 dict_FL AMPHETAMINES = {1: 'Oui', 0: 'Non'}
245 dict_FL METHAMPHETAMINES = {1: 'Oui', 0: 'Non'}
246 dict_FL MDMA = {1: 'Oui', 0: 'Non'}
247 dict_FL MEPHEDRONE = {1: 'Oui', 0: 'Non'}
248 dict_FL_OTHER_STIMULANTS = {1: 'Oui', 0: 'Non'}
249 dict_FL HYPNOTICS = {1: 'Oui', 0: 'Non'}
250 dict_FL BARBITURATES = {1: 'Oui', 0: 'Non'}
251 dict_FL BENZODIAZEPINES = {1: 'Oui', 0: 'Non'}
252 dict_FL GHB = {1: 'Oui', 0: 'Non'}
253 dict_FL_OTHER_HYPNOTICS = {1: 'Oui', 0: 'Non'}
254 dict_FL HALLUCINOGENS = {1: 'Oui', 0: 'Non'}
255 dict_FL LSD = {1: 'Oui', 0: 'Non'}
256 dict_FL KETAMINE = {1: 'Oui', 0: 'Non'}
257 dict_FL_OTHER_HALLUCINOGENS = {1: 'Oui', 0: 'Non'}
258 dict_FL VOLATILE = {1: 'Oui', 0: 'Non'}
259 dict_FL CANNABIS = {1: 'Oui', 0: 'Non'}
260 dict_FL MARIJUANA = {1: 'Oui', 0: 'Non'}
261 dict_FL HASH = {1: 'Oui', 0: 'Non'}
262 dict_FL_OTHER_CANNABIS = {1: 'Oui', 0: 'Non'}
263 dict_FL ALCOHOL = {1: 'Oui', 0: 'Non'}
264 dict_FL_OTHER_SUBSTANCE = {1: 'Oui', 0: 'Non'}
265 dict_CD_INJECTING_STAT = {1: 'Oui', 2: 'Non'}
266
267 # Function that allows recoding of the different variables
268
269 def funct_recode(variable):
270     if variable == 'CD_SEX':
271         return dict_sex[variable]
272     elif variable == 'CD_LVN_STAT_WHERE':
273         return dict_lvn_where[variable]
274     elif variable == 'CD_LVN_STAT_WITH_WHOM':
275         return dict_lvn_with_whom[variable]
276     elif variable == 'CD_LVN_CHILDREN':
277         return dict_lvn_children[variable]
278     elif variable == 'CD_EDUCATION_LEVEL':
279         return dict_educLevel[variable]
280     elif variable == 'CD_LABOUR_STAT':
281         return dict_labour_stat[variable]
282     elif variable == 'CD_INCOME_STAT':
283         return dict_income_stat[variable]
284     if variable == 'CD_REFERRAL':
285         return dict_referral[variable]
286     elif variable == 'CD_PREV_TREAT':
287         return dict_prev_treat[variable]
288     elif variable == 'CD_ROUTE_ADMIN':
289         return dict_route_admin[variable]
290     elif variable == 'CD_FREQ_USE':

```

```

291     return dict_freq_use[variable]
292 elif variable == 'FL_OPIATES':
293     return dict_FL_OPIATES[variable]
294 elif variable == 'FL_HEROIN':
295     return dict_FL_HEROIN[variable]
296 elif variable == 'FL_BUPRENORPHINE':
297     return dict_FL_BUPRENORPHINE[variable]
298 elif variable == 'FL_METHADONE':
299     return dict_FL_METHADONE[variable]
300 if variable == 'FL_FENTANYL':
301     return dict_FL_FENTANYL[variable]
302 elif variable == 'FL_OTHER_OPIATES':
303     return dict_FL_OTHER_OPIATES[variable]
304 elif variable == 'FL_COCAINE':
305     return dict_FL_COCAINE[variable]
306 elif variable == 'FL_POWDER_COCAINE':
307     return dict_FL_POWDER_COCAINE[variable]
308 elif variable == 'FL_CRACK':
309     return dict_FL_CRACK[variable]
310 elif variable == 'FL_OTHER_COCAINE':
311     return dict_FL_OTHER_COCAINE[variable]
312 elif variable == 'FL_STIMULANTS':
313     return dict_FL_STIMULANTS[variable]
314 elif variable == 'FL_AMPHETAMINES':
315     return dict_FL_AMPHETAMINES[variable]
316 elif variable == 'FL_METHAMPHETAMINES':
317     return dict_FL_METHAMPHETAMINES[variable]
318 elif variable == 'FL_MDMA':
319     return dict_FL_MDMA[variable]
320 elif variable == 'FL_MEPHEDRONE':
321     return dict_FL_MEPHEDRONE[variable]
322 elif variable == 'FL_OTHER_STIMULANTS':
323     return dict_FL_OTHER_STIMULANTS[variable]
324 elif variable == 'FL_HYPNOTICS':
325     return dict_FL_HYPNOTICS[variable]
326 elif variable == 'FL_BARBITURATES':
327     return dict_FL_BARBITURATES[variable]
328 elif variable == 'FL_BENZODIAZEPINES':
329     return dict_FL_BENZODIAZEPINES[variable]
330 elif variable == 'FL_GHB':
331     return dict_FL_GHB[variable]
332 elif variable == 'FL_OTHER_HYPNOTICS':
333     return dict_FL_OTHER_HYPNOTICS[variable]
334 elif variable == 'FL_HALLUCINOGENS':
335     return dict_FL_HALLUCINOGENS[variable]
336 elif variable == 'FL_KETAMINE':
337     return dict_FL_KETAMINE[variable]
338 elif variable == 'FL_OTHER_HALLUCINOGENS':
339     return dict_FL_OTHER_HALLUCINOGENS[variable]

```

```

340     elif variable == 'FL_LSD':
341         return dict_FL_LSD[variable]
342     elif variable == 'FL_VOLATILE':
343         return dict_FL_VOLATILE[variable]
344     elif variable == 'FL_CANNABIS':
345         return dict_FL_CANNABIS[variable]
346     elif variable == 'FL_MARIJUANA':
347         return dict_FL_MARIJUANA[variable]
348     elif variable == 'FL_HASH':
349         return dict_FL_HASH[variable]
350     elif variable == 'FL_OTHER_CANNABIS':
351         return dict_FL_OTHER_CANNABIS[variable]
352     elif variable == 'FL_ALCOHOL':
353         return dict_FL_ALCOHOL[variable]
354     elif variable == 'FL_OTHER_SUBSTANCE':
355         return dict_FL_OTHER_SUBSTANCE[variable]
356     elif variable == 'CD_INJECTING_STAT':
357         return dict_CD_INJECTING_STAT[variable]
358
359 # Define the data dictionary
360 data_dict = {'CD_SEX': dict_sex,
361             #'CD_REG': dict_reg,
362             'CD_EDUCATION_LEVEL': dict_educLevel,
363             'CD_LVN_STAT_WHERE': dict_lvn_where,
364             'CD_LVN_STAT_WITH_WHOM': dict_lvn_whith_whom,
365             'CD_LVN_CHILDREN': dict_lvn_children,
366             'CD_LABOUR_STAT': dict_labour_stat,
367             'CD_INCOME_STAT': dict_incomme_stat,
368             'CD_REFERRAL': dict_referral,
369             'CD_ROUTE_ADMIN': dict_route_admin,
370             'CD_FREQ_USE': dict_freq_use,
371             'CD_PREV_TREAT': dict_prev_treat,
372             'FL_OPIATES': dict_FL_OPIATES,
373             'FL HEROIN': dict_FL HEROIN,
374             'FL BUPRENORPHINE': dict_FL BUPRENORPHINE,
375             'FL METHADONE': dict_FL METHADONE,
376             'FL FENTANYL': dict_FL FENTANYL,
377             'FL_OTHER_OPIATES': dict_FL_OTHER_OPIATES,
378             'FL COCAINE': dict_FL COCAINE,
379             'FL POWDER COCAINE': dict_FL POWDER COCAINE,
380             'FL CRACK': dict_FL CRACK,
381             'FL_OTHER_COCAINE': dict_FL_OTHER_COCAINE,
382             'FL STIMULANTS': dict_FL STIMULANTS,
383             'FL AMPHETAMINES': dict_FL AMPHETAMINES,
384             'FL METHAMPHETAMINES': dict_FL METHAMPHETAMINES,
385             'FL MDMA': dict_FL MDMA,
386             'FL MEPHEDRONE': dict_FL MEPHEDRONE,
387             'FL_OTHER_STIMULANTS': dict_FL_OTHER_STIMULANTS,
388             'FL HYPNOTICS': dict_FL HYPNOTICS,

```

```

389         'FL_BARBITURATES': dict_FL_BARBITURATES,
390         'FL_BENZODIAZEPINES': dict_FL_BENZODIAZEPINES,
391         'FL_GHB': dict_FL_GHB,
392         'FL_OTHER_HYPNOTICS': dict_FL_OTHER_HYPNOTICS,
393         'FL_HALLUCINOGENS': dict_FL_HALLUCINOGENS,
394         'FL_LSD': dict_FL_LSD,
395         'FL_KETAMINE': dict_FL_KETAMINE,
396         'FL_OTHER_HALLUCINOGENS':
397             dict_FL_OTHER_HALLUCINOGENS,
398         'FL_VOLATILE': dict_FL_VOLATILE,
399         'FL_CANNABIS': dict_FL_CANNABIS,
400         'FL_MARIJUANA': dict_FL_MARIJUANA,
401         'FL_HASH': dict_FL_HASH,
402         'FL_OTHER_CANNABIS': dict_FL_OTHER_CANNABIS,
403         'FL_ALCOHOL': dict_FL_ALCOHOL,
404         'FL_OTHER_SUBSTANCE': dict_FL_OTHER_SUBSTANCE,
405         'CD_INJECTING_STAT': dict_CD_INJECTING_STAT}
406
407 # Redefine variable labels
408 for var, label_dict in data_dict.items():
409     df_cannabis_interest[var].replace(label_dict, inplace=True)
410
411 df_cannabis_interest['FL_CANNABIS'].unique()    #
412 A single category (of course)
413
414 # Deletion of the variable "FL_CANNABIS" (as it contains only one
415     category)
416 df_cannabis_interest.drop(['FL_CANNABIS'], axis=1, inplace=True)
417
418 ## Extraction of patients who initiated treatment in 2021 and
419     have never received treatment before.
420 # Converting the column "NR_DATE_START_TREAT" to type datetime64[
421     ns]
422 df_cannabis_interest['NR_DATE_START_TREAT'] = pd.to_datetime(
423     df_cannabis_interest['NR_DATE_START_TREAT'])
424
425 # Extraction of patients who initiated treatment in 2021 and have
426     never received treatment before
427 df_cannabis_2021 = df_cannabis_interest[(df_cannabis_interest['
428     NR_DATE_START_TREAT'].dt.year == 2021) &
429     (df_cannabis_interest['
430         CD_PREV_TREAT'] == 'Non
431         ')]
432
433 df_cannabis_2021.head(1)
434
435 # Resetting the indexes and deleting the 'CD_PREV_TREAT' variable
436     and the date
437 df_cannabis_2021 = df_cannabis_2021.reset_index(drop=True)
438 df_cannabis_2021.drop(['CD_PREV_TREAT'], axis=1, inplace=True)

```

```

428 df_cannabis_2021.drop(['NR_DATE_START_TREAT'], axis=1, inplace=
    True)
429 df_cannabis_2021.drop(['CD_ROUTE_ADMIN'], axis=1, inplace= True)
430 df_cannabis_2021.head(1)
431
432 # Number of substances uses per patient
433
434 df_cann2021 = df_cannabis_2021.copy()
435
436 variables = ['opiates','cocaine', 'stimulants_autresCocaine', '
    hypnotiques', 'hallucinogenes', 'inhalants_volatiles', '
    cannabis', 'alcool', 'autre_substance']
437 for col in variables:
438     df_cann2021[col] = 0
439
440 # This function calculates the number of substances uses per
    patient
441 def compute_sum(row):
442     dic = {'opiates': ['FL_HEROIN', 'FL_BUPRENORPHINE', 'FL_METHADONE
        ', 'FL_FENTANYL', 'FL_OTHER_OPIATES'],
443           'cocaine': ['FL_POWDER_COCAINE', 'FL_CRACK', '
        FL_OTHER_COCAINE'],
444           'stimulants_autresCocaine': {'FL_STIMULANTS', '
        FL_AMPHETAMINES', 'FL_METHAMPHETAMINES', 'FL_MDMA', '
        FL_MEPHEDRONE', 'FL_OTHER_STIMULANTS'},
445           'hypnotiques': ['FL_BARBITURATES', 'FL_BENZODIAZEPINES', '
        FL_GHB', 'FL_OTHER_HYPNOTICS'],
446           'hallucinogenes': ['FL_LSD', 'FL_KETAMINE', '
        FL_OTHER_HALLUCINOGENS'],
447           'inhalants_volatiles': ['FL_VOLATILE'],
448           'cannabis': ['FL_MARIJUANA', 'FL_HASH', 'FL_OTHER_CANNABIS'
        ],
449           'alcool': {'FL_ALCOHOL'},
450           'autre_substance' : {'FL_OTHER_SUBSTANCE'}}
451     ret = []
452     for var in dic:
453         retvar = 0
454         for ele in dic[var]:
455             a = 0
456             if row[ele] == 'Oui':
457                 a= 1
458             retvar = retvar + a
459         ret.append(retvar)
460     return ret
461
462 df_cann2021['opiates'], df_cann2021['cocaine'], df_cann2021['
    stimulants_autresCocaine'], df_cann2021['hypnotiques'],
    df_cann2021['hallucinogenes'], df_cann2021['inhalants_volatiles'
    ], df_cann2021['cannabis'], df_cann2021['alcool'], df_cann2021[

```

```

    'autre_substance'] = zip(*df_cann2021.apply(compute_sum, axis
    = 1))
463
464 df_cann2021['Total_subst_consom'] = 0
465 for col in variables:
466     df_cann2021['Total_subst_consom'] = df_cann2021['
    Total_subst_consom'] + df_cann2021[col]
467 df_cann2021.head()
468
469 # Transformation of the total number of substances consumed into
    a categorical variable
470 def recode_consomTot(col):
471     if col <= 2:
472         return '1-2'
473     elif col >= 6:
474         return '6-10'
475     else:
476         return ' 3-5'
477
478 df_cann2021['nombre_subst_total'] = df_cann2021['
    Total_subst_consom'].map(recode_consomTot)
479 df_cann2021['nombre_subst_total'].unique()
480
481 # Saving the database as a CSV file
482 df_cann2021.to_csv('df_cann2021.csv', index = False)
483
484
485 ## Data exploration
486
487 cols = ['NR_AGE', 'NR_DATE_START_TREAT', 'MS_AGE_FIRST_USE']
488 for col in list(df_cannabis_2021.columns):
489     if col not in cols:
490         df_cannabis_2021[col] = df_cannabis_2021[col].astype('object'
    )
491
492 # The following functions can be used to construct strip diagrams
    for a given variable
493
494 def plot_barplot_with_hue(dataframe, x,y, hue, title='', xlabel='
    ', ylabel='Fréquence'):
495
496     df=dataframe.groupby([x, hue])[y].count().reset_index()
497     df['pourcentage'] = df.groupby([x]).transform(lambda x: (x/x.
    sum()))*100)
498     df.fillna(0, inplace=True)
499     df['pourcentage'] = np.round(df['pourcentage'],1)
500     # plot
501     fig, ax = plt.subplots (figsize=(20, 10))

```

```

502 plots = sns.barplot(x=x, y='Fréquence', data = df, hue=hue, ec=
    "k",
503                     #order = df_kmodes[var].value_counts().
                        reset_index()['index'].tolist(),
504                     linewidth=1, palette="crest")
505 tailles = df[x].value_counts().tolist()
506 for bar in plots.patches:
507     plots.annotate(format(bar.get_height(), '.1f'),
508                   (bar.get_x() + bar.get_width() / 2,
509                    bar.get_height()), ha='center', va='center',
510                    size=15, xytext=(0, 8),
511                    textcoords='offset points')
512
513 plt.title(title, fontweight = 'bold', fontsize= 15)
514 ax.set_xlabel(xlabel)
515 ax.set_ylabel(ylabel)
516 # ax.set_xticklabels(ax.get_xticklabels(), rotation=10, ha="
    right")
517 ax.set_yticks([0,10,20,30,40,50,60,70,80,90,100])
518 plt.show()
519
520 def plot_barplot_without_hue(dataframe, x, title='', xlabel='',
    ylabel='Fréquence'):
521
522     df = dataframe[x].value_counts().rename_axis(x).reset_index(
        name='count')
523     df['pourcentage'] = (df['count'] / df['count'].sum()) * 100
524     df.fillna(0, inplace=True)
525     df['Fréquence'] = np.round(df['pourcentage'],1)
526     # plot
527     fig, ax = plt.subplots(figsize=(15, 10))
528     plots = sns.barplot(x=x, y='Fréquence', data = df, ec="k",
529                        linewidth=1, palette="crest")
530     tailles = df[x].value_counts().tolist()
531     for bar in plots.patches:
532         plots.annotate(format(bar.get_height(), '.1f'),
533                       (bar.get_x() + bar.get_width() / 2,
534                        bar.get_height()), ha='center', va='center',
535                       ,
536                       size=15, xytext=(0, 8),
537                       textcoords='offset points')
538     plt.title(title, fontweight = 'bold', fontsize= 15)
539     if x in ['CD_REFERRAL', 'CD_FREQ']:
540         ax.set_xticklabels(ax.get_xticklabels(), rotation=10, ha="
            right")
541         ax.set_yticks([0,5,10,15,20,25,30,35,40,45,50,55,60])
542         ax.set_xlabel(xlabel)
543         ax.set_ylabel(ylabel)
544         plt.show()

```

```

544
545 # Commented out IPython magic to ensure Python compatibility.
546 # %config InlineBackend.print_figure_kwargs={'facecolor' : "w"}
547 plt.style.use('classic')
548
549 # Descriptive statistics of quantitative variables
550 df_cannabis_2021[df_cannabis_2021.select_dtypes('float').columns
551                  ].describe().style\
552                  .set_table_styles([{"selector": "th", "props": "background-
553                                     color:darkcyan; font-size:1.5em; font-weight:bold;"},
554                                     {"selector": "td:hover", "props": "
555                                     background-color:darkcyan"},
556                                     {"selector": "td", "props": "border:1px
557                                     solid darkorchid; text-align:center;"}])
558
559 d = rp.summarize(df_cannabis_2021[["NR_AGE", "MS_AGE_FIRST_USE"
560                                   ]], decimals =2,
561                  stats= ["Mean","Median",'Min', 'Max',"Variance",
562                           "SD", "SE"])
563 (df_cannabis_2021[df_cannabis_2021.select_dtypes('float').columns
564                   ]).corr() # corrélation faible : résultat prévisible
565
566 fig, ax = plt.subplots(1, 1, figsize=(7, 8))
567 plt.subplots_adjust(top=1., right=2.5, wspace=.5, hspace=.5)
568 colors = np.repeat(["blueviolet", "goldenrod"], 1)
569 for i, col in enumerate(df_cannabis_2021.select_dtypes('float').
570                          columns):
571     sns.kdeplot(x=col, data=df_cannabis_2021, fill=True, legend=
572                 True)
573 ax.legend = ['c','f']
574 plt.show()
575 sns.set_style('darkgrid')
576 sns.set_context('paper',font_scale=1.5)
577 plt.style.use('ggplot')
578
579 # Histogram of ages
580
581 #fig,ax = plt.subplots(figsize=(15,15))
582 sns.histplot(df_cannabis_2021['NR_AGE'], kde=True, color = '
583              darkcyan', ec='k')
584 plt.title('Histogramme des âges')
585 ax.set_xlabel('NR_AGE')
586 plt.ylabel('Nombre de patients')
587 plt.axis('auto')
588 plt.xticks(np.arange(0, 70, 5))
589 plt.show()
590
591 # Histogram of ages at first cannabis use

```

```

582 sns.histplot(df_cannabis_2021['MS_AGE_FIRST_USE'], kde=True ,
    color = 'darkcyan', ec='k')
583 #plt.title('Histogramme des âges lors du premier usage du
    cannabis')
584 plt.title('Histogramme des âges lors du premier usage du cannabis
    ')
585 ax.set_xlabel('âges')
586 plt.ylabel('Nombre de patients')
587 plt.xticks(np.arange(5, 60, 5))
588 plt.show()
589
590 # age distribution by gender
591 fig = plt.figure(figsize=(10, 6))
592 sns.boxplot(x='CD_SEX', y='NR_AGE', data=df_cannabis_2021)
593 #plt.legend(loc='upper left')
594 plt.yticks(np.arange(5, 70, 5))
595 plt.show()
596
597 # Distribution of ages at first cannabis use by gender
598 plt.figure(figsize=(10, 6))
599 sns.boxplot(x='CD_SEX', y='MS_AGE_FIRST_USE', data=
    df_cannabis_2021)
600 plt.yticks(np.arange(5, 70, 5))
601 plt.show()
602
603
604 ##### Analysis of sociodemographic characteristics and types of
    substances used
605
606 # plot
607 for col in df_cannabis_2021.select_dtypes(include=['category']):
608     plot_barplot_without_hue(df_cannabis_2021.select_dtypes(
        include=['category']),col, xlabel=col)
609
610 # plot
611 for col in df_cannabis_2021.select_dtypes(include=['category']):
612     plot_barplot_without_hue(df_cannabis_2021.select_dtypes(
        include=['category']), col, xlabel=col)
613
614 features = ['CD_REFERRAL', 'CD_FREQ_USE', 'CD_INJECTING_STAT', '
    FL_OPIATES', 'FL_HEROIN',
615     'FL_BUPRENORPHINE', 'FL_METHADONE', 'FL_FENTANYL', '
    FL_OTHER_OPIATES',
616     'FL_COCAINE', 'FL_POWDER_COCAINE', 'FL_CRACK', '
    FL_OTHER_COCAINE',
617     'FL_STIMULANTS', 'FL_AMPHETAMINES', 'FL_METHAMPHETAMINES',
    'FL_MDMA',
618     'FL_MEPHEDRONE', 'FL_OTHER_STIMULANTS', 'FL_HYPNOTICS',

```

```

619         'FL_BARBITURATES', 'FL_BENZODIAZEPINES', 'FL_GHB', '
        FL_OTHER_HYPNOTICS',
620         'FL_HALLUCINOGENS', 'FL_LSD', 'FL_KETAMINE', '
        FL_OTHER_HALLUCINOGENS',
621         'FL_VOLATILE', 'FL_MARIJUANA', 'FL_HASH', '
        FL_OTHER_CANNABIS',
622         'FL_ALCOHOL', 'FL_OTHER_SUBSTANCE']
623
624 for var in features:
625     plot_barplot_without_hue(df_cannabis_2021, x=var, xlabel=var)
626
627 # Using the function "plot_barplot_without_hue" to construct the
    histograms
628 plot_barplot_without_hue(df_cann2021, x='nombre_subst_total',
    xlabel='nombre de substances problématiques (par catégories)')
629
630 ## Outlier treatment
631
632 # This function allows determining outliers using the Z-score
633 def z_score_outlier_detect(data, attr, thresh=3):
634     """
635     Function to determine outliers based on Z-Score
636
637     Parameters:
638     -----
639     data : pandas.DataFrame
640         Represents the data with which we are working.
641     attr : str
642         Represents the attribute name
643     thresh : int
644         Represents the threshold values (in this case the number
        of stnd deviations from the mean)
645     Returns:
646     -----
647     out : array-like
648         Represents the indices of outlier points in the data.
649     """
650     mu = data[attr].mean()
651     std = data[attr].std()
652
653     z_score = (data[attr] - mu) / std
654     out = np.where(np.abs(z_score) > thresh)[0]
655     return out
656
657 outlier_stats_z_score = list()
658
659 # Percentage of outliers per variable
660 for col in df_cannabis_2021[df_cannabis_2021.select_dtypes('float
    ')].columns:

```

```

661     _out = z_score_outlier_detect(df_cannabis_2021[
        df_cannabis_2021.select_dtypes('float').columns], col)
662     outlier_stats_z_score.append((col, len(_out) / df[col].shape
        [0] * 100))
663 df_outliers = pd.DataFrame.from_records(outlier_stats_z_score,
        columns=["attributes", "percent_outliers"])\
664 .set_index("attributes")\
665 .sort_values(by="percent_outliers").plot(kind="bar", figsize=(12,
        8),
666                                     ylabel="pourcentage d'
                                     outliers", xlabel = '
                                     variables', color="
                                     darkcyan", legend=
                                     False)
667 plt.show()
668
669 ## Application of mixed data factor analysis
670
671 #
672 This function applies MFA (Mixed Factor Analysis) on the data.
        The number of components is chosen to achieve 100% explained
        variance
673 def perform_famd(df, n_components):
674     famd = prince.FAMD(
675         n_components=n_components,
676         n_iter=3,
677         copy=True,
678         check_input=True,
679         random_state=42,
680         engine="sklearn"
681     )
682     famd = famd.fit(df)
683     return famd
684
685 # Application of the function "perform_famd" to the data
686 famd = perform_famd(df_cannabis_2021, 70)
687
688 # Function that gives the percentage of variance explained as a
        function of the number of components chosen
689
690 def total_var_expli(famd, n_components):
691     cumulative_var_ratio = famd.explained_inertia_.cumsum()
692     variance_explained = cumulative_var_ratio[n_components - 1] *
        100
693     return variance_explained
694
695 # Percentage of explained variance
696 variance_expliquee = total_var_expli(famd, 70)
697

```

```

698 # Coordinates of individuals in the principal component space
699
700 df_indiv_famd = famd.row_coordinates(df_cannabis_2021)
701 df_cannabis_2021_famd = df_indiv_famd.add_prefix('comp')
702 df_cannabis_2021_famd.head(2)
703
704 # Saving the result in a dataframe
705 df_var_famd = famd.column_coordinates_.add_prefix('comp')
706 df_var_famd.head(1)
707
708 # visualization
709 plt.figure(figsize=(15, 10))
710 famd.plot(
711     df_cannabis_2021,
712     x_component=0,
713     y_component=1
714 )
715
716 # Saving the result in a CSV file
717 df_cannabis_2021_famd.to_csv('df_cannabis_2021famd.csv', index =
    False)
718
719 # Contribution of variables to the formation of factor axes
720 famd.column_contributions_.add_prefix('comp').head(2)
721 len(famd.feature_names_in_)
722
723
724 #Initial 20 derived features contribution to the proposed model
725 plt.figure(figsize=(15, 10))
726 x= [i for i in range(20)]
727 plt.xticks(x)
728 plt.bar(x,famd.percentage_of_variance_[0:20], color = 'darkcyan')
729
730 # Clustering
731
732 # Dataframe to use in clustering: the clusters obtained from each
    method will be added to this dataframe.
733 df_can_2021_withClusters = df_cannabis_2021.copy()
734
735 # Saving to a dataframe
736 df_can2021_temp = df_cannabis_2021_famd.copy()
737 df_can2021_temp.head(2)
738
739 ### K-means
740
741 #### Selection of the optimal number of clusters
742
743 class OptimalK:
744     def __init__(self):

```

```

745     pass
746     # This function allows constructing the silhouette score as a
       function of the number of clusters and the intra-cluster
       inertia using the elbow method
747     def find_k(self, data, k_range):
748
749         # Using Elbow Method
750         cost = list()
751
752         # Using Silhouette's Method
753         sil = list()
754         for i in tqdm(k_range):
755             model = KMeans(n_clusters=i, random_state=42)
756             model.fit(data)
757             cost.append(model.inertia_)
758             labels = model.labels_
759             sil.append(silhouette_score(data, labels, metric='
               euclidean', sample_size=1000))
760
761         plt.figure(figsize=(15, 10))
762         plt.plot(k_range, cost, 'bx-', label="Elbow Method",
               color="royalblue")
763         plt.ylabel('Distortion')
764         ax.set_xlabel('number clusters')
765         plt.show()
766
767         plt.figure(figsize=(15, 10))
768         plt.plot(k_range, sil, 'bx-', label="Silhouette Method",
               color="darkcyan")
769         plt.ylabel('silhouette score')
770         ax.set_xlabel('number clusters')
771         plt.show()
772
773     op = OptimalK()
774     op.find_k(df_can2021_temp, range(2, 10))
775
776     # Calculating the silhouette score as a function of the number of
       clusters
777     print(__doc__)
778     data_ = df_can2021_temp.copy()
779
780     range_n_clusters = range(2, 9)
781
782     for n_clusters in range_n_clusters:
783         fig, ax = plt.subplots(1, 1)
784
785         ax.set_xlim([-0.5, 0.8])
786         ax.set_ylim([0, len(data_) + (n_clusters) * 10])
787         ax.set_yticks([])

```

```

788
789 clusterer = KMeans(n_clusters=n_clusters, random_state=10)
790 cluster_labels = clusterer.fit_predict(data_)
791
792 silhouette_avg = silhouette_score(data_, cluster_labels)
793 print("For n_clusters =", n_clusters, "The average
      silhouette_score is :", silhouette_avg)
794
795 sample_silhouette_values = silhouette_samples(data_,
      cluster_labels)
796
797 y_lower = 10
798 for i in range(n_clusters):
799     ith_cluster_silhouette_values = sample_silhouette_values[
      cluster_labels == i]
800     ith_cluster_silhouette_values.sort()
801
802     size_cluster_i = ith_cluster_silhouette_values.shape[0]
803     y_upper = y_lower + size_cluster_i
804
805     cmap = cm.get_cmap("Spectral")
806     color = cmap(float(i) / n_clusters)
807     ax.fill_betweenx(np.arange(y_lower, y_upper), 0,
      ith_cluster_silhouette_values,
808                     facecolor=color, edgecolor=color, alpha
      =1)
809
810     ax.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))
811
812     y_lower = y_upper + 10
813 plt.xlabel("Scores de silhouette")
814 plt.ylabel("clusters")
815 plt.axvline(x=silhouette_avg, color="red", linestyle="--")
816 plt.xticks(np.arange(-0.5, 0.8, 0.2))
817 plt.show()
818
819 ##### Visualization of the results from clustering into 2 clusters
820
821 model = KMeans(n_clusters=2, random_state=42)
822 model.fit(df_can2021_temp)
823 pred = model.predict(df_can2021_temp)
824
825
826 # save predictions in the dataframe
827 df_cannabis_2021_famd["clusters_kmeans_2"] = pred.astype(int)
828 df_cannabis_2021_famd["clusters_kmeans_2"] =
      df_cannabis_2021_famd["clusters_kmeans_2"].map({0: 'cluster_1',
829

```

```

830 df_can_2021_withClusters["clusters_kmeans_2"] =
      df_cannabis_2021_famd["clusters_kmeans_2"]
831
832 # visualization of partition into 2 clusters
833 plt.figure(figsize=(15, 10))
834 sns.scatterplot(x="comp0", y="comp1", hue="clusters_kmeans_2",
835               data=df_cannabis_2021_famd,
836               palette=dict(cluster_2="blue", cluster_1="green"),
837               hue_order = ["cluster_1", "cluster_2"])
838 plt.show()
839
840 # Number of patients per cluster
841 np.unique(df_cannabis_2021_famd["clusters_kmeans_2"],
842         return_counts=True)
843
844 # plot
845 ax = sns.countplot(x="clusters_kmeans_2", data=
846                 df_can_2021_withClusters,
847                 palette=dict(cluster_2="blue", cluster_1="green"),
848                 order=['cluster_1', 'cluster_2'])
849 ax.set_ylabel('Count', title='')
850 # add annotations
851 for c in ax.containers:
852
853     # custom label calculates percent and add an empty string so
854     # 0 value bars don't have a number
855     labels = [f'{h/df_can_2021_withClusters.clusters_kmeans_2.
856               count()*100:0.1f}%' if (h := v.get_height()) > 0 else ''
857               for v in c]
858     #plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
859     ax.bar_label(c, labels=labels, label_type='edge')
860 ax.set_xlabel('clusters')
861 plt.show()
862
863 ##### Visualization of the results from clustering into 3 clusters
864
865 kmeans_3 = KMeans(n_clusters=3, random_state=42)
866 kmeans_3.fit(df_can2021_temp)
867
868 pred_kmeans_3 = kmeans_3.predict(df_can2021_temp)
869
870 np.unique(pred_kmeans_3, return_counts=True)
871
872 # 3 clusters

```

```

867 kmeans_3test = KMeans(n_clusters=4, random_state=42)
868 kmeans_3test.fit(df_can2021_temp)
869
870 pred_kmeans_3test = kmeans_3test.predict(df_can2021_temp)
871
872 np.unique(pred_kmeans_3test, return_counts=True)
873
874 df_cannabis_2021_famd["clusters_kmeans_3"] = pred_kmeans_3.astype
      (int)
875
876 df_cannabis_2021_famd["clusters_kmeans_3"] =
      df_cannabis_2021_famd["clusters_kmeans_3"].map({0: 'cluster_1',
      1: 'cluster_2', 2: 'cluster_3'})
877
878 df_can_2021_withClusters["clusters_kmeans_3"] =
      df_cannabis_2021_famd["clusters_kmeans_3"]
879
880 np.unique(df_cannabis_2021_famd["clusters_kmeans_3"],
      return_counts=True)
881
882 # plot
883 ax = sns.countplot(x="clusters_kmeans_3", data=
      df_can_2021_withClusters,
884                  palette=dict(cluster_3="gold", cluster_2="blue",
      cluster_1="green"),
885                  order=['cluster_1', 'cluster_2', 'cluster_3'])
886 ax.set(ylabel='Count', title='')
887 # add annotations
888 for c in ax.containers:
889
890     # custom label calculates percent and add an empty string so
      0 value bars don't have a number
891     labels = [f'{h/df_can_2021_withClusters.clusters_kmeans_3.
      count()*100:0.1f}%' if (h := v.get_height()) > 0 else ''
      for v in c]
892     #plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
893     ax.bar_label(c, labels=labels, label_type='edge')
894 ax.set_xlabel('clusters')
895 plt.show()
896
897 plt.figure(figsize=(15, 10))
898 sns.scatterplot(x="comp0", y="comp1", hue="clusters_kmeans_3",
      data=df_cannabis_2021_famd, alpha=1,
899               palette=dict(cluster_3="gold", cluster_2="blue",
      cluster_1="green"),
900               hue_order = ["cluster_1", "cluster_2", "cluster_3"
      ])
901 plt.legend(title='Kmeans clusters')
902 plt.show()

```

```

903
904
905 ### Hierarchical Agglomerative Clustering
906
907 ##### Selection of the optimal number of clusters: dendrogram
908
909 # Construction of the dendrogram
910 Z = linkage(df_can2021_temp, 'ward')
911 plt.figure(figsize=(15, 10))
912 plt.title('Hierarchical Clustering Dendrogram')
913 plt.xlabel('')
914 plt.ylabel('distance de Ward')
915 dendrogram(
916     Z,
917     leaf_rotation=90.,
918     leaf_font_size=8.,
919 )
920 plt.show()
921
922
923 ##### Silhouette coefficient as a function of the number of
924     clusters
925
926 for n_clusters in range(2, 11):
927     fig, ax = plt.subplots(1, 1)
928     X = df_can2021_temp.copy()
929     ax.set_xlim([-0.5, 0.8])
930     ax.set_ylim([0, len(X) + (n_clusters) * 10])
931     ax.set_yticks([])
932
933     clusterer = AgglomerativeClustering(n_clusters=n_clusters,
934         affinity='euclidean', linkage='ward')
935     cluster_labels = clusterer.fit_predict(X)
936
937     silhouette_avg = silhouette_score(X, cluster_labels)
938     print("For n_clusters =", n_clusters, "The average
939         silhouette_score is :", silhouette_avg)
940
941     sample_silhouette_values = silhouette_samples(X,
942         cluster_labels)
943
944     y_lower = 10
945     for i in range(n_clusters):
946         ith_cluster_silhouette_values = sample_silhouette_values[
947             cluster_labels == i]
948         ith_cluster_silhouette_values.sort()
949
950         size_cluster_i = ith_cluster_silhouette_values.shape[0]
951         y_upper = y_lower + size_cluster_i

```

```

947     cmap = cm.get_cmap("Spectral")
948     color = cmap(float(i) / n_clusters)
949     plt.fill_betweenx(np.arange(y_lower, y_upper), 0,
950                       ith_cluster_silhouette_values,
951                       facecolor=color, edgecolor=color, alpha
952                           =0.7)
953     plt.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))
954     y_lower = y_upper + 10
955
956     plt.xticks(np.arange(-0.5, 0.8, 0.2))
957     plt.xlabel("Scores de silhouette")
958     plt.ylabel("Clusters")
959
960     plt.axvline(x=silhouette_avg, color="red", linestyle="--")
961 plt.show()
962
963 ##### Grouping into 2 clusters
964
965 hc_2 = AgglomerativeClustering(n_clusters=2, affinity='euclidean',
966                               linkage='ward')
967 hc_2.fit_predict(df_can2021_temp)
968
969 # saving clusters in the dataframe
970 df_cannabis_2021_famd["clusters_hc_2"] = hc_2.labels_
971 df_cannabis_2021_famd["clusters_hc_2"] = df_cannabis_2021_famd["
972     clusters_hc_2"].map({0: 'cluster_1', 1: 'cluster_2'})
973 df_can_2021_withClusters["clusters_hc_2"] = df_cannabis_2021_famd
974     ["clusters_hc_2"]
975
976 # Number of patients in each cluster
977 np.unique(df_cannabis_2021_famd["clusters_hc_2"], return_counts=
978     True)
979
980 # Visualization of the number of patients per cluster
981 ax = sns.countplot(x="clusters_hc_2", data=
982     df_can_2021_withClusters,
983     palette=dict(cluster_2="blue", cluster_1="green"),
984     order=['cluster_1', 'cluster_2'])
985 ax.set(ylabel='Count', title='')
986
987 # add annotations
988 for c in ax.containers:
989     # custom label calculates percent and add an empty string so
990     # 0 value bars don't have a number
991     labels = [f'{h/df_can_2021_withClusters.clusters_kmeans_3.
992         count()*100:0.1f}%' if (h := v.get_height()) > 0 else ''
993         for v in c]
994     plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
995     ax.bar_label(c, labels=labels, label_type='edge')

```

```

986 ax.set_xlabel('clusters')
987 plt.show()
988
989 plt.figure(figsize=(15, 10))
990
991 sns.scatterplot(x="comp0", y="comp1", hue="clusters_hc_2", data=
    df_cannabis_2021_famd,
992                 palette=dict(cluster_3="gold", cluster_2="blue",
    cluster_1="green"),
993                 hue_order = ["cluster_1", "cluster_2"])
994 plt.legend(title='HC clusters')
995 plt.show()
996
997 ### Description of the clusters
998
999 fig, ax = plt.subplots(figsize=(10, 6))
1000 sns.boxplot(x='clusters_hc_2', y='NR_AGE', data=
    df_can_2021_withClusters,
1001            hue_order = ["cluster_1", "cluster_2"],
1002            palette=dict(cluster_1="green", cluster_2="blue"))
1003 #plt.legend(loc='best')
1004 ax.set_xlabel('Sexe')
1005 ax.set_ylabel('Ages')
1006 ax.set_yticks([0,5,10,15,20,25,30,35,40,45,50,55,60,65])
1007
1008
1009 fig, ax = plt.subplots(figsize=(10, 6))
1010 sns.boxplot(x='clusters_hc_2', y='MS_AGE_FIRST_USE', data=
    df_can_2021_withClusters,
1011            #sns.boxplot(x='CD_SEX', y='MS_AGE_FIRST_USE', hue='
    clusters_hc_2', data=df_can_2021_withClusters,
1012            hue_order = ["cluster_1", "cluster_2"],
1013            palette=dict(cluster_1="green", cluster_2="blue"))
1014 ax.set_xlabel('Sexe')
1015 ax.set_ylabel('Ages à la première consommation')
1016 ax.set_yticks([0,5,10,15,20,25,30,35,40,45,50,55])
1017
1018 # Saving the database to a CSV file
1019 df_hc = df_can_2021_withClusters.drop(['clusters_kmeans_2', '
    clusters_kmeans_3'], axis= 1)
1020 df_hc.to_csv('df_hc.csv', index = False)
1021
1022 ### K-prototypes
1023 cols = ['NR_AGE', 'NR_DATE_START_TREAT', 'MS_AGE_FIRST_USE']
1024 for col in list(df_cannabis_2021.columns):
1025     if col not in cols:
1026         df_cannabis_2021[col] = df_cannabis_2021[col].astype('
    category')
1027

```

```

1028 #Pre-processing numerical data
1029 scaler = StandardScaler()
1030 df_scaled=df_cannabis_2021.copy()
1031 for c in list(df_cannabis_2021.select_dtypes(exclude=['category',
1032     ])).columns):
1033     df_scaled.loc[:,c] = scaler.fit_transform(np.array(
1034         df_cannabis_2021[c]).reshape(-1, 1)).flatten()
1035
1036 # select categorical data
1037 list_temp = list(range(0,43))
1038 list_temp.remove(1)
1039 list_temp.remove(40)
1040 categorical_columns = list_temp
1041 len(df_scaled.columns)
1042 len(categorical_columns)
1043 df_scaled.head(2)
1044
1045 ##### Selection of the optimal number of clusters
1046
1047 # Elbow technique
1048 costs = []
1049 n_clusters = []
1050 clusters_assigned = []
1051 for i in range(2, 8):
1052     #try:
1053     kproto = KPrototypes(n_clusters=i, random_state=10, init='
1054         Huang', n_init=1, verbose=1)
1055     clusters = kproto.fit_predict(df_scaled, categorical=
1056         categorical_columns)
1057     costs.append(kproto.cost_)
1058     n_clusters.append(i)
1059     clusters_assigned.append(clusters)
1060
1061 # visualization
1062 plt.figure(figsize=(15, 10))
1063 plt.plot(n_clusters, costs, 'bx-')
1064 plt.ylabel('Distortion')
1065 ax.set_xlabel('number clusters');
1066
1067 # Similarity measurement
1068 def mixed_distance(a,b,categorical=None, alpha=0.01):
1069     if categorical is None:
1070         num_score=kprototypes.euclidean_dissim(a,b)
1071         return num_score
1072     else:
1073         cat_index=categorical
1074         a_cat=[]
1075         b_cat=[]
1076         for index in cat_index:

```

```

1073         a_cat.append(a[index])
1074         b_cat.append(b[index])
1075     a_num=[]
1076     b_num=[]
1077     l=len(a)
1078     for index in range(l):
1079         if index not in cat_index:
1080             a_num.append(a[index])
1081             b_num.append(b[index])
1082
1083     a_cat=np.array(a_cat).reshape(1,-1)
1084     a_num=np.array(a_num).reshape(1,-1)
1085     b_cat=np.array(b_cat).reshape(1,-1)
1086     b_num=np.array(b_num).reshape(1,-1)
1087     cat_score=kprototypes.matching_dissim(a_cat,b_cat)
1088     num_score=kprototypes.euclidean_dissim(a_num,b_num)
1089     return cat_score+num_score*alpha
1090 def dm_prototypes(dataset,categorical=None,alpha=0.8):
1091     #if the input dataset is a dataframe, take out the values as
1092     #a numpy.
1093     #If the input dataset is a numpy array, use it as is.
1094     if type(dataset).__name__=='DataFrame':
1095         dataset=dataset.values
1096     lenDataset=len(dataset)
1097     distance_matrix=np.zeros(lenDataset*lenDataset).reshape(
1098         lenDataset,lenDataset)
1099     for i in range(lenDataset):
1100         for j in range(lenDataset):
1101             x1= dataset[i]
1102             x2= dataset[j]
1103             distance=mixed_distance(x1, x2,categorical=
1104                 categorical_columns,alpha=alpha)
1105             distance_matrix[i][j]=distance
1106             distance_matrix[j][i]=distance
1107     return distance_matrix
1108
1109 # Attention! This code takes approximately 20 minutes to execute
1110 silhouette_scores = dict()
1111 K = range(2,10)
1112 distance_matrix=dm_prototypes(df_scaled,categorical=
1113     categorical_columns,alpha=0.8)
1114 for k in K:
1115     untrained_model = kprototypes.KPrototypes(n_clusters=k,
1116         max_iter=20)
1117     trained_model = untrained_model.fit(df_scaled, categorical=
1118         categorical_columns)
1119     cluster_labels = trained_model.labels_
1120     score=silhouette_score(distance_matrix, cluster_labels,metric
1121         ="precomputed")

```

```

1115     silhouette_scores[k]=score
1116
1117 ##### Calculating the silhouette score as a function of the number
        of clusters
1118 for n_clusters in range(2,9):
1119     fig, ax = plt.subplots(1, 1)
1120     X=df_scaled.copy()
1121     ax.set_xlim([-0.5, 0.8])
1122     ax.set_ylim([0, len(X) + (n_clusters) * 10])
1123     ax.set_yticks([])
1124     clusterer = kprototypes.KPrototypes(n_clusters=n_clusters,
        random_state=10,max_iter=20)
1125     cluster_labels = clusterer.fit_predict(X, categorical=
        categorical_columns)
1126
1127     silhouette_avg = silhouette_score(distance_matrix,
        cluster_labels)
1128     print("For n_clusters =", n_clusters,
1129           "The average silhouette_score is :", silhouette_avg)
1130     # Compute the silhouette scores for each sample
1131     sample_silhouette_values = silhouette_samples(distance_matrix
        , cluster_labels)
1132     y_lower = 10
1133     for i in range(n_clusters):
1134         ith_cluster_silhouette_values = sample_silhouette_values[
            cluster_labels == i]
1135         ith_cluster_silhouette_values.sort()
1136
1137         size_cluster_i = ith_cluster_silhouette_values.shape[0]
1138         y_upper = y_lower + size_cluster_i
1139
1140         cmap = cm.get_cmap("Spectral")
1141         color = cmap(float(i) / n_clusters)
1142
1143         plt.fill_betweenx(np.arange(y_lower, y_upper), 0,
            ith_cluster_silhouette_values,
1144                           facecolor=color, edgecolor=color, alpha
            =0.7)
1145         plt.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))
1146         y_lower = y_upper + 10
1147
1148     plt.xticks(np.arange(-0.5, 0.8, 0.2))
1149     plt.xlabel("Scores de silhouette")
1150     plt.ylabel("Clusters")
1151     plt.axvline(x=silhouette_avg, color="red", linestyle="--")
1152 plt.show()
1153
1154 ##### Visualization of the predictions (partition into 2 clusters)

```

```

1155 kproto = KPrototypes(n_clusters=2, random_state=10, init='Huang',
1156                       n_init=1, verbose=1)
1157
1158 clusters = kproto.fit_predict(df_scaled, categorical=
1159                               categorical_columns)
1160
1161 # Adding the clusters to the database
1162 df_cannabis_2021_famd['clusters_kproto'] = kproto.labels_
1163 df_can_2021_withClusters['Cluster_kproto'] = kproto.labels_
1164 df_cannabis_2021_famd["clusters_kproto"] = df_cannabis_2021_famd[
1165     "clusters_kproto"].map({0: 'cluster_1',
1166
1167
1168
1169
1170
1171
1172 df_can_2021_withClusters["clusters_kproto"] =
1173     df_cannabis_2021_famd["clusters_kproto"]
1174
1175 # Visualization of the clusters
1176 plt.figure(figsize=(15, 10))
1177 sns.scatterplot(x="comp0", y="comp1", hue="clusters_kproto", data
1178                =df_cannabis_2021_famd,
1179                palette=dict(cluster_2="blue", cluster_1="green")
1180                )
1181 plt.legend(title='Kprototypes clusters')
1182 plt.show()
1183
1184 # Number of patients in each cluster.
1185 np.unique(df_cannabis_2021_famd["clusters_kproto"], return_counts
1186         =True)
1187
1188 # plot
1189 ax = sns.countplot(x="clusters_kproto", data=
1190                   df_can_2021_withClusters,
1191                   palette=dict(cluster_2="blue", cluster_1="green"),
1192                   order=['cluster_1', 'cluster_2'])
1193 ax.set(ylabel='Count', title='')
1194 # add annotations
1195 for c in ax.containers:
1196
1197     # custom label calculates percent and add an empty string so
1198     # 0 value bars don't have a number
1199     labels = [f'{h/df_can_2021_withClusters.clusters_kmeans_3.
1200                count()*100:0.1f}%' if (h := v.get_height()) > 0 else ''
1201               for v in c]
1202     #plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
1203     ax.bar_label(c, labels=labels, label_type='edge')
1204 #plt.xlabel('clusters')

```

```

1188 plt.show()
1189
1190 df_can_2021_withClusters['CD_INCOME_STAT'].value_counts().
    reset_index()
1191 df_can_2021_withClusters['CD_INCOME_STAT'].value_counts().
    reset_index()['index'].tolist()
1192 df_can_2021_withClusters['CD_INCOME_STAT'].value_counts().
    reset_index()['CD_INCOME_STAT'].tolist()
1193 df_can_2021_withClusters['CD_INCOME_STAT'].value_counts().tolist
    ()
1194 features = ['CD_SEX', 'CD_EDUCATION_LEVEL', 'CD_LABOUR_STAT', '
    CD_INCOME_STAT', 'CD_LVN_STAT_WITH_WHOM', 'CD_LVN_CHILDREN']
1195
1196 for var in features:
1197     plt.figure(figsize=(15,8))
1198     ax = sns.countplot(x=var,data=df_can_2021_withClusters, hue=
        'clusters_kproto',
1199                             palette=dict(cluster_2="blue", cluster_1="
                                green", cluster_3 = 'gold'),
1200                             hue_order = ["cluster_1","cluster_2"],ec="k
                                ",
1201                             order = df_can_2021_withClusters[var].
                                value_counts().reset_index()['index'].
                                tolist(),
1202                             linewidth=1,saturation = 1)
1203     #plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
1204     ax.set_xticklabels(ax.get_xticklabels(), rotation=10, ha="
        right")
1205     tailles = df_can_2021_withClusters[var].value_counts().tolist
        ()
1206
1207     for c in ax.containers:
1208         labels = [f'{{(v.get_height())/(tailles[k])*100:0.1f}}' if (v
            .get_height()) > 0 else '' for k,v in enumerate(c)]
1209         ax.bar_label(c, labels=labels, label_type='edge')
1210     plt.legend(title='Kprototypes clusters')
1211     plt.show()
1212
1213 # Types of substances consumed
1214 features = ['CD_REFERRAL', 'CD_FREQ_USE', 'FL_ALCOHOL', 'FL HEROIN
    ',
1215             'FL_COCAINE', 'FL_MARIJUANA', 'FL_CRACK', 'FL_KETAMINE',
            'FL_HALLUCINOGENS']
1216
1217 for var in features:
1218     plt.figure(figsize=(15,8))
1219     ax = sns.countplot(x=var,data=df_can_2021_withClusters, hue=
        'clusters_kproto',

```

```

1220         palette=dict(cluster_2="blue", cluster_1="
1221             green"),
1222         hue_order = ["cluster_1", "cluster_2"], ec="k
1223             ",
1224         order = df_can_2021_withClusters[var].
1225             value_counts().reset_index()['index'].
1226             tolist(),
1227         linewidth=1, saturation = 1)
1228 #plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
1229 ax.set_xticklabels(ax.get_xticklabels(), rotation=10, ha="
1230     right")
1231 tailles = df_can_2021_withClusters[var].value_counts().tolist
1232     ()
1233
1234 for c in ax.containers:
1235     labels = [f' {(v.get_height())/(tailles[k])*100:0.1f}' if (v
1236         .get_height()) > 0 else '' for k,v in enumerate(c)]
1237     ax.bar_label(c, labels=labels, label_type='edge')
1238 plt.legend(title='Kprototypes clusters')
1239 plt.show()
1240
1241 fig = plt.figure(figsize=(10, 6))
1242 sns.boxplot(x='CD_SEX', y='NR_AGE', hue='clusters_kproto', data=
1243     df_can_2021_withClusters,
1244     palette=dict(cluster_2="blue", cluster_1="green"))
1245 #plt.legend(loc='upper right', bbox_to_anchor=(0, 0))
1246 #plt.legend(title='Kprototypes clusters')
1247
1248 fig = plt.figure(figsize=(10, 6))
1249 sns.boxplot(x='CD_SEX', y='MS_AGE_FIRST_USE', hue='
1250     clusters_kproto', data=df_can_2021_withClusters,
1251     palette=dict(cluster_2="blue", cluster_1="green"))
1252 #plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
1253 plt.legend(title='Kproto clusters')
1254
1255 ## K-modes
1256
1257 # Converting age into a categorical variable
1258 df_kmodes = df_cannabis_2021.copy()
1259 df_kmodes['age_bin'] = pd.cut(df_cannabis_2021['NR_AGE'], [10,
1260     15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65],
1261     labels=['10-15 ans', '15-20 ans', '
1262         20-25 ans', '25-30 ans', '30-35 ans
1263         ', '35-40 ans', '40-45 ans', '45-50
1264         ans', '50-55 ans', '55-60 ans', '
1265         60-65 ans'])
1266 df_kmodes = df_kmodes.drop('NR_AGE', axis = 1)

```

```

1255 # Transformation of age at first consumption
1256 df_kmodes['AGE_FIRST_bin'] = pd.cut(df_cannabis_2021['
1257     MS_AGE_FIRST_USE'], [5, 10, 15, 20,25, 30,35, 40,45, 50,55],
1258     labels=['5-10 ans', '10-15 ans', '
1259         15-20 ans', '20-25 ans', '25-30 ans
1260         ', '30-35 ans', '35-40 ans', '40-45
1261         ans', '45-50 ans', '50-55 ans'])
1262 df_kmodes = df_kmodes.drop('MS_AGE_FIRST_USE',axis = 1)
1263 df_kmodes['AGE_FIRST_bin'].unique()
1264 df_kmodes['age_bin'].value_counts().reset_index()['index'].tolist
1265 ()
1266 df_kmodes['age_bin'].value_counts().tolist()
1267 # Visualization of the distribution of ages by categories
1268 plot_barplot_without_hue(df_kmodes, x='age_bin', xlabel= "âges
1269     par catégories")
1270 plot_barplot_without_hue(df_kmodes, x='AGE_FIRST_bin', xlabel= "â
1271     ges lors du premier usage du cannabis (par catégories)")
1272 ##### Selection of the optimal number of clusters using the
1273     silhouette score
1274 # We perform label encoding to encode the variables
1275 le = preprocessing.LabelEncoder()
1276 df_cannabis_new = df_kmodes.apply(le.fit_transform)
1277 df_cannabis_new.head(2)
1278 cost = []
1279 for num_clusters in list(range(2,10)):
1280     kmode = KModes(n_clusters=num_clusters, init = "Cao", n_init
1281         = 1, verbose=1)
1282     kmode.fit_predict(df_cannabis_new)
1283     cost.append(kmode.cost_)
1284 # Elbow method
1285 y = np.array([i for i in range(2,10,1)])
1286 plt.figure(figsize=(15, 10))
1287 plt.plot(y, cost, 'bx-', label="Elbow Method", color="blue")
1288 plt.ylabel('Distortion')
1289 #plt.xlabel('number clusters')
1290 plt.show()
1291 km_cao = KModes(n_clusters=3, init = "Cao", n_init = 1, verbose
1292     =1)
1293 fitClusters_cao = km_cao.fit_predict(df_cannabis_new)
1294 # Using the silhouette coefficient

```

```

1294 print(__doc__)
1295 range_n_clusters = range(2, 9)
1296 for n_clusters in range_n_clusters:
1297     fig, ax = plt.subplots(1, 1)
1298
1299     # The 1st subplot is the silhouette plot
1300     # The silhouette coefficient can range from -1, 1 but in this
1301     # example all
1302     # lie within [-0.5, 0.8]
1303     ax.set_xlim([-0.5, 0.8])
1304     ax.set_ylim([0, len(data_) + (n_clusters) * 10])
1305     ax.set_yticks([])
1306
1307     # Initialize the clusterer with n_clusters value and a random
1308     # generator
1309     # seed of 10 for reproducibility.
1310     clusterer = KModes(n_clusters=n_clusters, init = "Cao",
1311                        n_init = 1, verbose=1)
1312     cluster_labels = clusterer.fit_predict(df_cannabis_new)
1313
1314     # The silhouette_score gives the average value for all the
1315     # samples.
1316     # This gives a perspective into the density and separation of
1317     # the formed
1318     # clusters
1319     silhouette_avg = silhouette_score(df_cannabis_new,
1320                                     cluster_labels)
1321     print("For n_clusters =", n_clusters,
1322           "The average silhouette_score is :", silhouette_avg)
1323
1324     # Compute the silhouette scores for each sample
1325     sample_silhouette_values = silhouette_samples(df_cannabis_new
1326                                                  , cluster_labels)
1327
1328     y_lower = 10
1329     for i in range(n_clusters):
1330         ith_cluster_silhouette_values = sample_silhouette_values[
1331             cluster_labels == i]
1332         ith_cluster_silhouette_values.sort()
1333
1334         size_cluster_i = ith_cluster_silhouette_values.shape[0]
1335         y_upper = y_lower + size_cluster_i
1336
1337         cmap = cm.get_cmap("Spectral")
1338         color = cmap(float(i) / n_clusters)
1339
1340         plt.fill_betweenx(np.arange(y_lower, y_upper), 0,
1341                           ith_cluster_silhouette_values,

```

```

1333         facecolor=color, edgecolor=color, alpha
1334             =0.7)
1335     plt.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))
1336     y_lower = y_upper + 10
1337
1338     plt.xticks(np.arange(-0.6, 0.8, 0.2))
1339     plt.xlabel("Scores de silhouette")
1340     plt.ylabel("Clusters")
1341
1342     plt.axvline(x=silhouette_avg, color="red", linestyle="--")
1343 plt.show()
1344
1345 # Adding the predictions to the database
1346 df_can_2021_withClusters['clusters_kmodes'] = fitClusters_cao
1347 df_cannabis_2021_famd['clusters_kmodes'] = fitClusters_cao
1348
1349 df_cannabis_2021_famd["clusters_kmodes"] = df_cannabis_2021_famd[
1350     "clusters_kmodes"].map({0: 'cluster_1', 1: 'cluster_2', 2: '
1351     cluster_3'})
1352
1353 df_can_2021_withClusters["clusters_kmodes"] =
1354     df_cannabis_2021_famd["clusters_kmodes"]
1355
1356 ##### Visualization of the results
1357
1358 np.unique(df_cannabis_2021_famd["clusters_kmodes"], return_counts
1359     =True)
1360
1361 # Visualisation du la partion en 3 clusters
1362 ax = sns.countplot(x="clusters_kmodes", data=
1363     df_can_2021_withClusters,
1364     palette=dict(cluster_3="gold", cluster_2="blue",
1365         cluster_1="green"),
1366     order=['cluster_1', 'cluster_2', 'cluster_3'])
1367 ax.set(ylabel='Count', title='')
1368 # add annotations
1369 for c in ax.containers:
1370
1371     # custom label calculates percent and add an empty string so
1372     # 0 value bars don't have a number
1373     labels = [f'{h/df_can_2021_withClusters.clusters_kmeans_3.
1374         count()*100:0.1f}%' if (h := v.get_height()) > 0 else ''
1375         for v in c]
1376     #plt.legend(loc='upper left', bbox_to_anchor=(1, 1))
1377     ax.bar_label(c, labels=labels, label_type='edge')
1378 ax.set_xlabel('clusters')
1379 plt.show()
1380
1381 plt.figure(figsize=(15, 10))

```

```

1372 sns.scatterplot(x="comp0", y="comp1", hue="clusters_kmodes", data
      =df_cannabis_2021_famd,
1373                  palette=dict(cluster_3="gold", cluster_2="blue",
                                cluster_1="green"),
1374                  hue_order = ["cluster_1", "cluster_2", "cluster_3"
                                ])
1375 plt.legend(title='Kmodes clusters')
1376 plt.show()
1377
1378 # save results in a .csv file
1379 df_can_2021_withClusters.to_csv('df_can_2021_withClusters.csv',
      index=False)
1380
1381 # Comparison between the K-means, Hierarchical Agglomerative
      Clustering (CAH), K-modes, and K-prototypes methods
1382
1383 df_compa = pd.DataFrame({'nombre_clusters': [2,3,4,5,6,7], 'K-
      means': [0.372,0.048,0.022,0.031,0.033,0.073], 'CAH'
      : [0.448,0.447,0.438,0.417,0.028,0.031], 'K-modes'
      : [0.040,0.073,-0.006,-0.033,-0.076,-0.043],
1384                  'K-prototypes': [0.324,0.299,0.215,0.148,
                                0.154, 0.138]})
1385
1386 sns.lineplot(data=df_compa, x='nombre_clusters', y='K-means',
      label='K-means')
1387 sns.lineplot(data=df_compa, x='nombre_clusters', y='CAH', label='
      CAH')
1388 sns.lineplot(data=df_compa, x='nombre_clusters', y='K-modes',
      label='K-modes')
1389 sns.lineplot(data=df_compa, x='nombre_clusters', y='K-prototypes'
      , label='K-prototypes')
1390
1391 plt.xlabel('Nombre de clusters')
1392 plt.ylabel('Score de silhouette moyen')
1393 plt.xticks(df_compa['nombre_clusters'])
1394 plt.ylim(-0.10, 0.50) # Définit l'intervalle des valeurs sur l'
      axe des ordonnées
1395
1396 plt.show()
1397
1398 sns.lineplot(data= df_compa, x='nombre_clusters', y='K-means',
      label='K-means')
1399 sns.lineplot(data= df_compa, x='nombre_clusters', y='CAH', label=
      'CAH')
1400 sns.lineplot(data= df_compa, x='nombre_clusters', y='K-modes',
      label='K-modes')
1401 sns.lineplot(data= df_compa, x='nombre_clusters', y='K-prototypes
      ', label='K-prototypes')
1402 plt.xticks([1,2,3,4,5,6,7,8])

```

```

1403 plt.yticks
      ([-0.10,-0.05,0.00,0.05,0.10,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.50])

1404 plt.xlabel('nombre de clusters')
1405 plt.ylabel('score de silhouette moyen')
1406
1407 df_nbreEnregis = pd.DataFrame({'années'
      : [2015,2016,2017,2018,2019,2020,2021], 'nbre de patients'
      : [2258,2191,2249,2171,2174,1741,1847]})
1408 df_nbreEnregis['date'] = pd.date_range(start='2015', end='2022',
      freq='Y')
1409 df_nbreEnregis.index = df_nbreEnregis.date
1410 df_nbreEnregis['nbre de patients'].plot()
1411 plt.xlim('2014-01-01', '2022-01-01')
1412 plt.yticks([0,250,500,750,1000,1250,1500,1750,2000,2250,2500])
1413 plt.xlabel('années')
1414 plt.ylabel('nombre de patients')
1415 plt.xticks(['2015-01-01', '2016-01-01', '2017-01-01', '2018-01-01',
      '2019-01-01', '2020-01-01', '2021-01-01'])
1416
1417 sns.lineplot(data= df_nbreEnregis, x='années', y='nbre de
      patients')
1418 plt.xticks([2014,2015,2016,2017,2018,2019,2020,2021,2022])
1419
1420 # Discussion: Analysis of the clusters
1421 # Data reading
1422
1423 # The following database contains all the individuals associated
      with their respective clusters. It can therefore be downloaded
      and used as required
1424
1425
1426
1427 df_can_2021_withClusters = pd.read_csv('/content/drive/MyDrive/
      memoire_keline/donne es/Segmentation study/Data/
      df_can_2021_withClusters.csv', sep=',')
1428
1429 selected_labels = df_can_2021_withClusters['clusters_hc_2'] #
      This can be modified depending on the chosen clustering
      approach.
1430 method_selected = 'clusters_hc_2' # Name of the selected
      variable in the table containing all the clustering results (
      df_can_2021_withClusters)
1431
1432 ##### Importance of variables in cluster formation
1433
1434 df_cat = df_cannabis_2021.select_dtypes(include=['category'])
1435 df_cat.head(1)
1436

```

```

1437 def chi_test(ds, labels):
1438     ct = pd.crosstab(ds, labels)
1439     return chi2_contingency(ct)
1440
1441 res_cat_variables = df_cat.apply(lambda x: chi_test(x,
1442     selected_labels)[-1]).T
1443
1444 res_cat_variables.columns = ["chi2", "p-valeur", "df"]
1445
1446
1447 res_cat_variables.sort_values("p-valeur")
1448
1449 # Identification of statistically significant variables (p-values
1450 < 0.05)
1451
1452 def apply_formatting(col):
1453     if col.name == 'p-valeur':
1454         return ['background-color: red' if c > 0.05 else '
1455             background-color: green' for c in col.values]
1456
1457 res = res_cat_variables.style.apply(apply_formatting, subset=['p-
1458     valeur'])
1459
1460 res
1461
1462 latex = res.render(formatter={ 'p-valeur': '{:.2f}'.format },
1463     escape=False)
1464
1465 with open('table.tex', 'w') as f:
1466     f.write(latex)
1467
1468
1469 ##### Characterization of the clusters
1470 ##### Visualization of the distribution of patient
1471     characteristics within the clusters
1472
1473 selected_labels.value_counts(ascending=True).reset_index()['index
1474     '].tolist()
1475
1476
1477 selected_labels.value_counts(ascending=True).tolist()
1478
1479
1480 ##### Analysis of sociodemographic characteristics
1481
1482 def compute_percents(dataframe, variable, clusters=
1483     method_selected, rowname=''):
1484     """
1485     This function calculates the contingency table between each
1486     categorical variable and the clusters, as well as the
1487     proportions in each cluster
1488 
```

```

1476     """
1477     crossTab = pd.crosstab(df_can_2021_withClusters[variable], [
1478         df_can_2021_withClusters[method_selected]],
1479         rownames=[rowname], colnames=['clusters'], margins=
1480             True)
1481     orange = sns.light_palette("green", as_cmap=True) # color
1482     proportions = pd.crosstab(index=df_can_2021_withClusters[
1483         variable], columns=df_can_2021_withClusters[method_selected],
1484         rownames=[rowname],
1485         colnames=['clusters'],
1486         margins=True, margins_name="TOTAL",
1487         normalize = 'columns').round(4).style.
1488         background_gradient(cmap=orange).format("{:.2%}")
1489     return crossTab, proportions
1490
1491 """"--- Sexe""""
1492
1493 crossTab, proportions = compute_percents(df_can_2021_withClusters
1494     , variable='CD_SEX')
1495 proportions
1496 crossTab
1497
1498 """"--- Source principale de revenus""""
1499
1500 crossTab, proportions = compute_percents(df_can_2021_withClusters
1501     , variable='CD_INCOME_STAT')
1502 proportions
1503 crossTab
1504
1505 crossTab, proportions = compute_percents(df_can_2021_withClusters
1506     , variable='CD_EDUCATION_LEVEL')
1507 proportions
1508 crossTab
1509
1510 crossTab, proportions = compute_percents(df_can_2021_withClusters
1511     , variable='CD_LVN_STAT_WHERE')
1512 proportions
1513 crossTab
1514
1515 crossTab, proportions = compute_percents(df_can_2021_withClusters
1516     , variable='CD_LVN_STAT_WITH_WHOM')
1517 proportions
1518 crossTab
1519
1520 crossTab, proportions = compute_percents(df_can_2021_withClusters
1521     , variable='CD_LVN_CHILDREN')
1522 proportions
1523 crossTab
1524
1525 crossTab, proportions = compute_percents(df_can_2021_withClusters
1526     , variable='CD LABOUR_STAT')

```

```

1514 proportions
1515 crossTab
1516
1517 ##### Analyse des âges
1518
1519 # analyse globale
1520 d = rp.summarize(df_can_2021_withClusters[["NR_AGE", "
1521     MS_AGE_FIRST_USE"]],
1522     stats= ["Mean", "Median", "Variance", "SD", 'Min',
1523         'Max', "Kurtosis", "Skew"])
1524
1525 # analyse des âges au sein des clusters
1526 d=rp.summarize(df_can_2021_withClusters.groupby([method_selected
1527     ])[["MS_AGE_FIRST_USE"], decimals = 2, stats= ["Mean", "Median", '
1528     Min', 'Max', "Variance", "SD", "SE", "CI"])
1529
1530 d
1531 ##### Analysis of the main reason for referral to treatment
1532 crossTab, proportions = compute_percents(df_can_2021_withClusters
1533     , variable='CD_REFERRAL')
1534
1535 d9 =rp.summarize(df_can_2021_withClusters.groupby([
1536     method_selected])[["NR_AGE", "MS_AGE_FIRST_USE"]], decimals =
1537     2, stats= ["N", "Mean", "Median", 'Min', 'Max', "Variance", "SD", "
1538     SE", "CI"])
1539
1540 """"##### Analysis of the types of substances used within the
1541     clusters""""
1542 crossTab, proportions = compute_percents(df_can_2021_withClusters
1543     , variable='CD_FREQ_USE')
1544
1545 # Opiacés
1546 crossTab, proportions = compute_percents(df_can_2021_withClusters
1547     , variable='FL_OPIATES')
1548
1549 crossTab, proportions = compute_percents(df_can_2021_withClusters
1550     , variable = 'FL_OTHER_OPIATES')
1551
1552 # Cocaïne
1553 crossTab, proportions = compute_percents(df_can_2021_withClusters
1554     , variable='FL_OTHER_STIMULANTS')
1555
1556 # Alcool
1557 crossTab, proportions = compute_percents(df_can_2021_withClusters
1558     , variable='FL_ALCOHOL')
1559
1560 # Cannabis
1561 crossTab, proportions = compute_percents(df_can_2021_withClusters
1562     , variable='FL_MARIJUANA')
1563
1564 crossTab, proportions = compute_percents(df_can_2021_withClusters
1565     , variable='FL_HASH')

```

```

1548
1549 # Hypnotiques ou sédatifs
1550 crossTab, proportions = compute_percents(df_can_2021_withClusters
    , variable='FL_OTHER_HYPNOTICS')
1551
1552
1553 # Hallucinogènes
1554
1555 crossTab, proportions = compute_percents(df_can_2021_withClusters
    , variable='FL_OPIATES')
1556
1557 # Autres substances
1558 crossTab, proportions = compute_percents(df_can_2021_withClusters
    , variable='FL_OTHER_SUBSTANCE')

```

