Université de Lille
1 SCIENCES ET TECHNOLOGIES

Mathématiques et Informatique Appliquées aux Science Humains et Sociales (MIASHS)

Option :

Méthodes quantitatives et modélisations pour l'entreprise (MQME)

Stage Master 1

Modéliser le suivi biologique des patients DNID et déterminer des sous-populations de patients à partir de leur observance du suivi biologique.

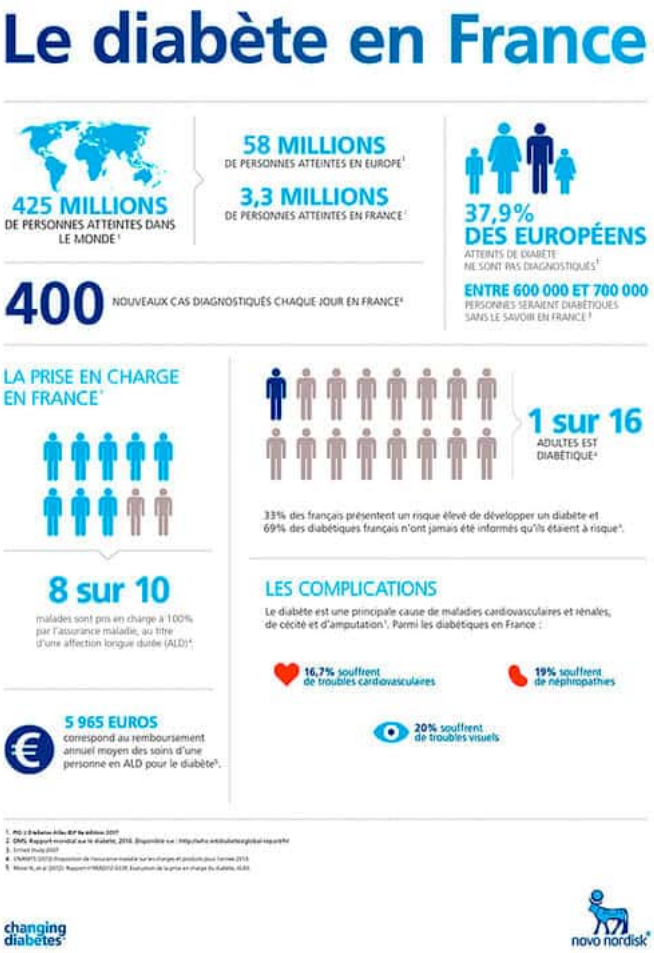
Travail réalisé par :
Ahmed ali HADAD
Mai 2024 - Juin 2024

Sous l'encadrement de :
Mme **Sophie Dabo** professeure chercheuse des université
Universdité de Lille 1

INTRODUCTION

Le diabète est une maladie insidieuse, touchant 3,5 millions de personnes en France et évolutive avec des complications dramatiques (cécité, maladie rénale, accidents vasculaires, artérite...) qui pourraient être évitables. D'après plusieurs études réalisées, l'observance des patients est proche de 40% et le niveau de suivi biologique est également très faible.

Un article publié en Novembre 2023, à regarde si besoin via le lien : Institut Amelis - Comprendre le diabète : types, causes, symptômes et traitements



Sous la supervision de l'enseignante-chercheuse **Madame Sophie Dabo**, nous cherchons à modéliser le suivi biologique des patients DNID du laboratoire **QuantiHealth** et de déterminer des sous-populations de patients (en utilisant des méthodes de partitionnement) à partir de leur observance du suivi biologique.

Pour modéliser le suivi biologique et déterminer d'avantage des sous-populations de patients, nous parcourons trois parties :

→ Etude des observance pour comprendre l'evolution, le suivi médicale, des patients.

→ Méthode de partitionnement : Classification Ascendante Hiérarchique (CAH).

→ Méthode de partitionnement : K-means.

Présentation du laboratoire

QuantiHealth est un laboratoire de recherche et d'analyse innovant spécialisé dans le domaine de la santé quantitative. Fort d'une expertise multidisciplinaire, QuantiHealth se distingue par son approche intégrative combinant des techniques avancées de biostatistique, de bioinformatique et d'épidémiologie pour améliorer la compréhension et la gestion des problèmes de santé.

Domaines d'expertise :

a. Biostatistique et Bioinformatique :

→ Analyse des données cliniques et génomiques pour identifier les biomarqueurs et les profils de risque.

→ Développement de modèles prédictifs pour la prévention, le diagnostic et le traitement des maladies.

→ Utilisation d'algorithmes de machine learning et d'intelligence artificielle pour exploiter de grands ensembles de données.

b. Épidémiologie et Santé Publique :

→ Études épidémiologiques pour comprendre la distribution et les déterminants des maladies dans les populations.

→ Conception et mise en œuvre d'enquêtes de santé publique pour informer les politiques de santé.

→ Suivi et analyse des tendances épidémiologiques pour prévoir les épidémies et les pandémies.

Services et Solutions :

a. Consultation et Support Technique :

→ Assistance aux chercheurs et cliniciens dans la conception d'études et l'analyse de données.

b. Analyse de Données :

→ Services d'analyse de données pour des projets de recherche académique et industrielle.

→ Plateformes d'analyse en ligne pour l'exploration des données génomiques et cliniques.

c. Recherche Collaborative :

→ Partenariats avec des institutions académiques, des hôpitaux et des entreprises pharmaceutiques pour des projets de recherche collaborative.

→ Participation à des consortiums internationaux et des initiatives de recherche.

Engagement Éthique et Développement Durable :

→ QuantiHealth adhère à des normes éthiques strictes pour la protection des données et la confidentialité des patients.Assistance aux chercheurs et cliniciens dans la conception d'études et l'analyse de données.

→ Engagement envers le développement durable et l'utilisation responsable des ressources.

Table des matières

1

Étude des observance des patients

Nous travaillons avec une base de données contenant des observations biologiques. Il est donc essentiel de mettre en lumière ces observations, de les décrire, de les étudier, de les analyser et de les modéliser.

Pour ce faire, nous avons généralement deux parties à aborder :

a. Présentation de la base de données

b. Étude descriptive des données

1.1 Présentation de la base de données

La base de données contient les observations de **55 426 patients distincts**. Elle comprend **844 170 observations** (lignes) et **19 colonnes**.

Une observation représente une consultation d'un patient, ce qui signifie que la base contient **844 170 consultations**. Par conséquent, un patient peut avoir une ou plusieurs consultations.

En ce qui concerne les colonnes, nous avons :

code patient	L'identifiant du patient qui est unique pour chaque patient.
date dossier	Le jour de la consultation.
mois de naissance	Mois de naissance du patient.
année de naissance	Année de naissance du patient.
sexe	Si le patient est mâle ou femelle.
code postal tronqué	Le code postal du patient.
HbA1c (%)	Le taux de concentration de l'hémoglobine glyquée dans le sang, qui est un reflet de la glycémie.
glycémie à jeun (mmol/L)	Le taux de concentration de la glycémie dans le sang.
créatininémie (mg/L)	Le taux de concentration de la créatinine dans l'urine.
créatininurie (mg/L)	Le taux de concentration de la créatinine dans le sang.
cholestérol total (g/L)	Le taux de concentration d cholestérol dans le sang.
HDL (g/L)	Le HDL (High-Density Lipoprotein), est le taux de fraction de cholestérol transportée par les lipoprotéines de haute densité dans le sang.
triglycéride (g/L)	Le taux de concentration triglycérides dans le sang, causé par la graisse.

microalbuminurie (mg/L)	Le taux de concentration d'albumine dans l'urine.
protéinurie (g/L)	Le taux de concentration de protéines dans l'urine.
RAC (mmol/L)	Le **RAC** (Rapport Albumine/Créatinine), le taux de concentration d'albumine dans l'urine par rapport à la quantité de créatinine.
protéinurie 24h (g/24h)	Le taux de concentration de protéines présentes dans l'urine collectée sur une période de 24 heures.
ASAT (UI/L)	ASAT (Aspartate Aminotransférase), l'acivité enzymatique des ASAT dans le corps.
ALAT (UI/L)	ALAT (Alanine Aminotransférase), l'acivité enzymatique des ASAT dans le métabolisme des acides aminés.

TABLE 1.1 – Description des variables

Généralement, un **diabète** est considéré comme **équilibré** ou **controlé**, c'est-à-dire ne représentant pas une menace, si le **taux d'HbA1c** est inférieur ou égal à 7%. Au-delà, le risque de développer des complications à long terme augmente. Une augmentation de 1% du **taux d'HbA1c** représente une augmentation moyenne de la glycémie de 0,30 g/l.

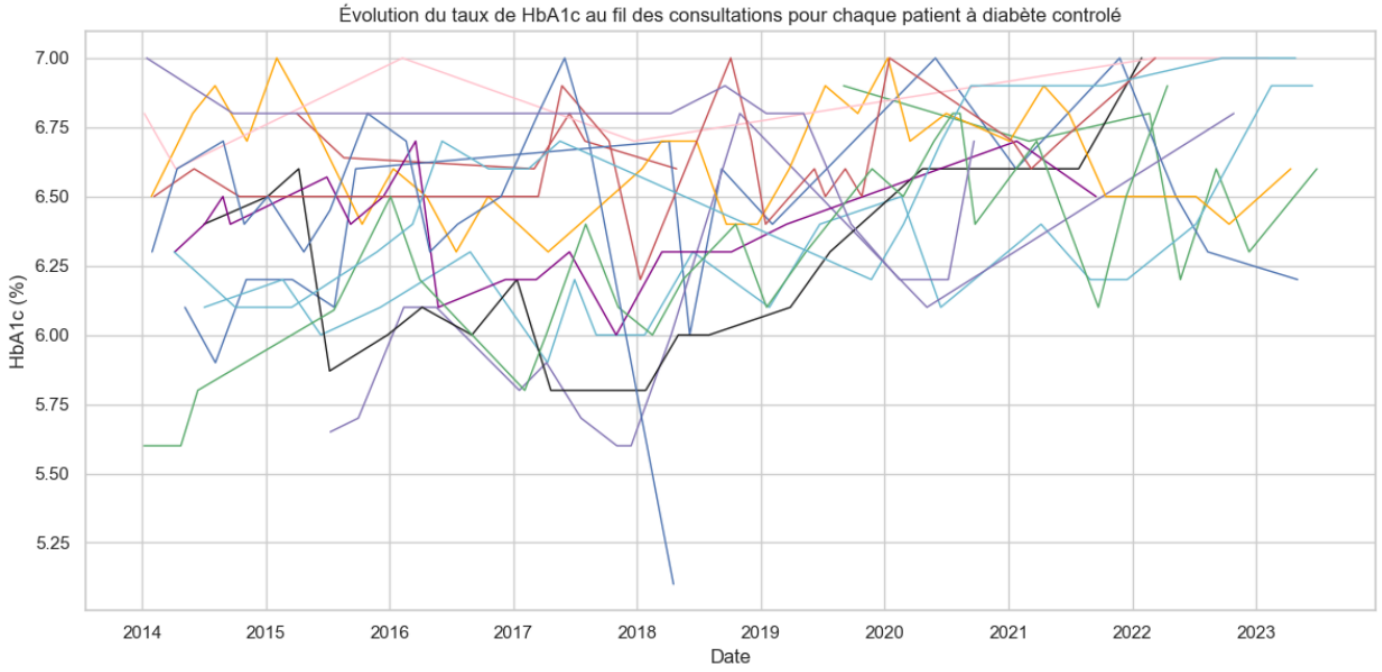
1.2 Etude de données

Dans cette partie, nous parcourons les données, les analysons et les nettoyons avant de procéder au partitionnement. Les points essentiels abordés sont les suivants :

1.2.1 Rytme de consultation

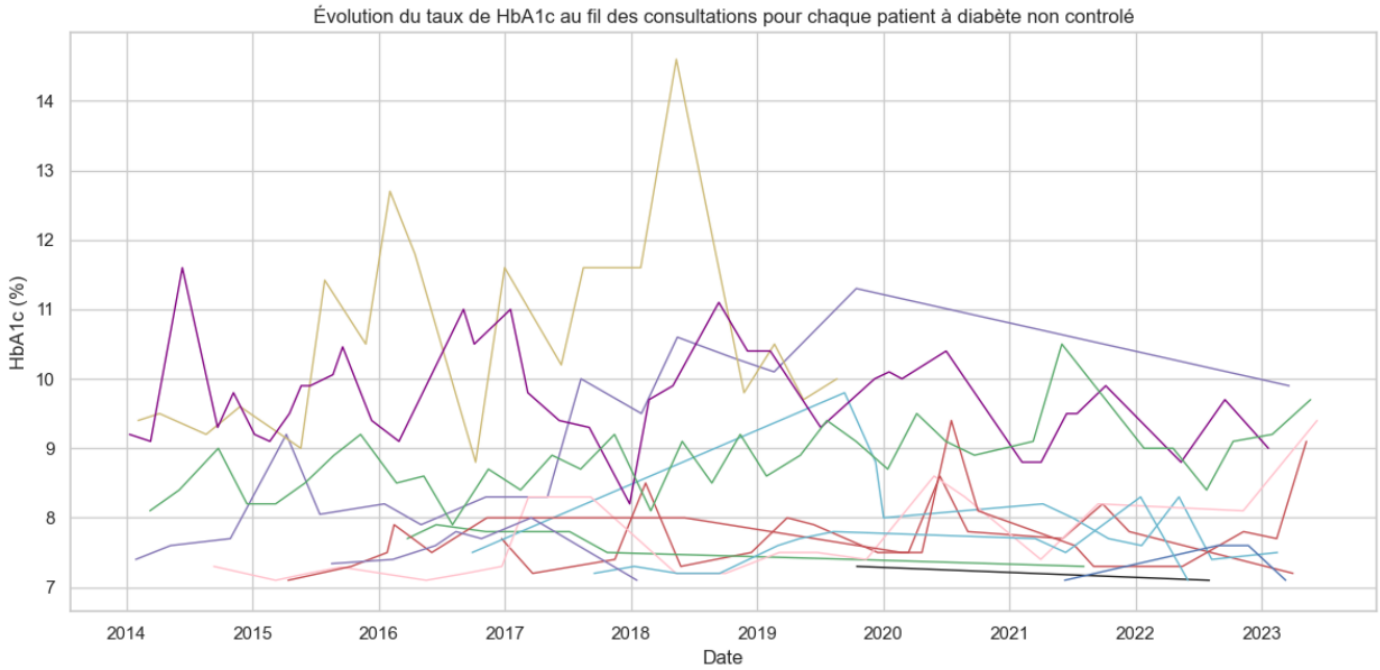
Analyse de la régularité du suivi médical et de l'évolution du taux d'HbA1c.

Il est important de noter que l'étude du diabète se concentre principalement sur la mesure du **taux d'HbA1c**. Nous allons donc examiner la régularité des consultations pour les patients atteints de diabète équilibré (contrôlé), c'est-à-dire ceux dont le taux d'HbA1c est inférieur ou égal à 7%, ainsi que pour les patients atteints de diabète non contrôlé.



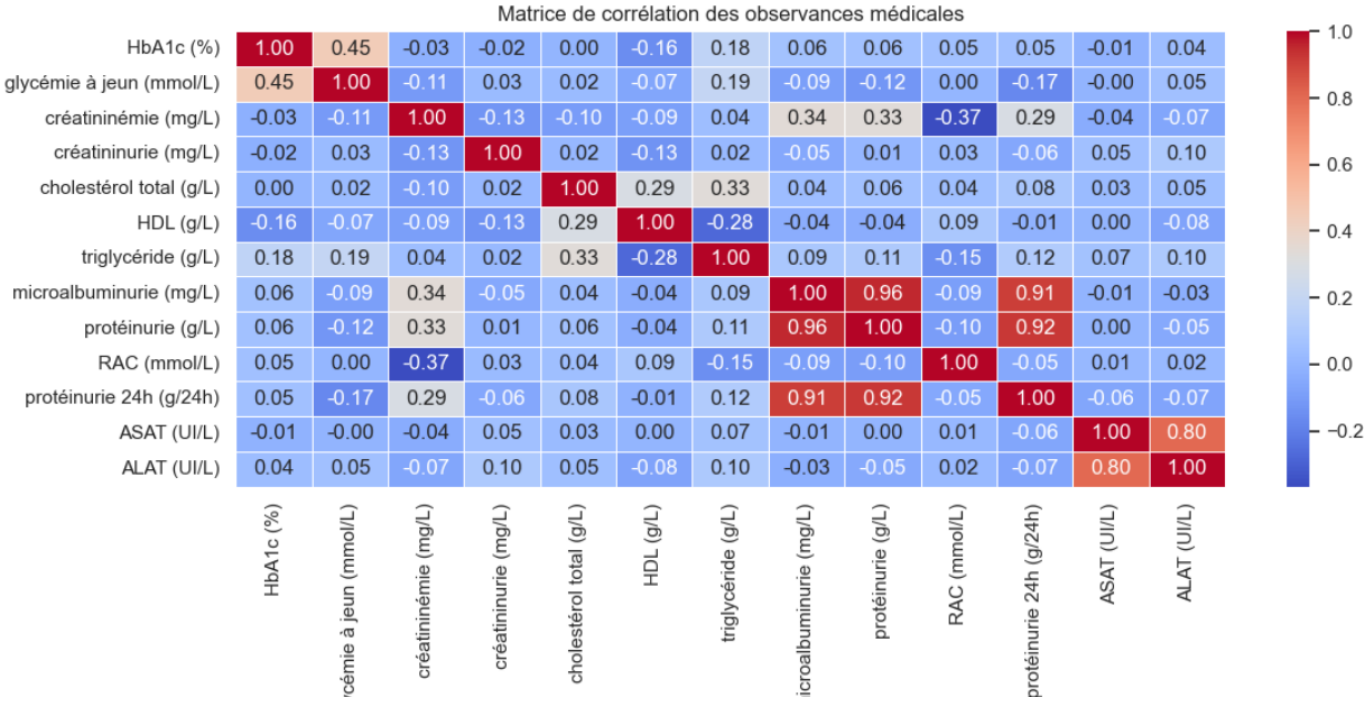
Cette figure présente l'évolution du taux d'HbA1c au fil du temps pour **15 patients** dont **le diabète est contrôlé**. On remarque que les pics de taux d'HbA1c sont assez réguliers, ce qui peut indiquer une régularité des consultations médicales. Cependant, il y a des cas où les consultations semblent être

moins fréquentes. Par exemple, pour un patient, la courbe est constante de 2015 à 2018, puis montre une augmentation continue de 2020 à 2023.



1.2.2 Etude des correlations

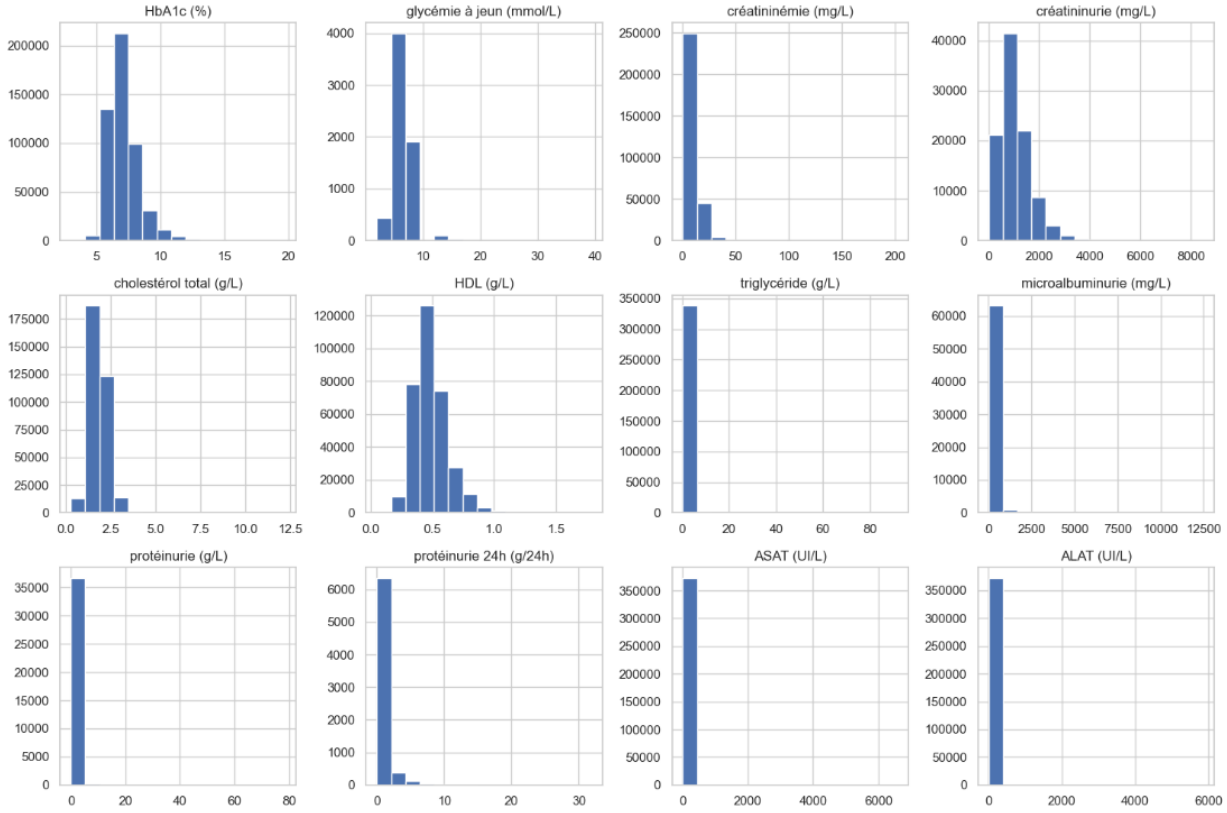
Exploration de la dépendance entre certaines observations médicales.



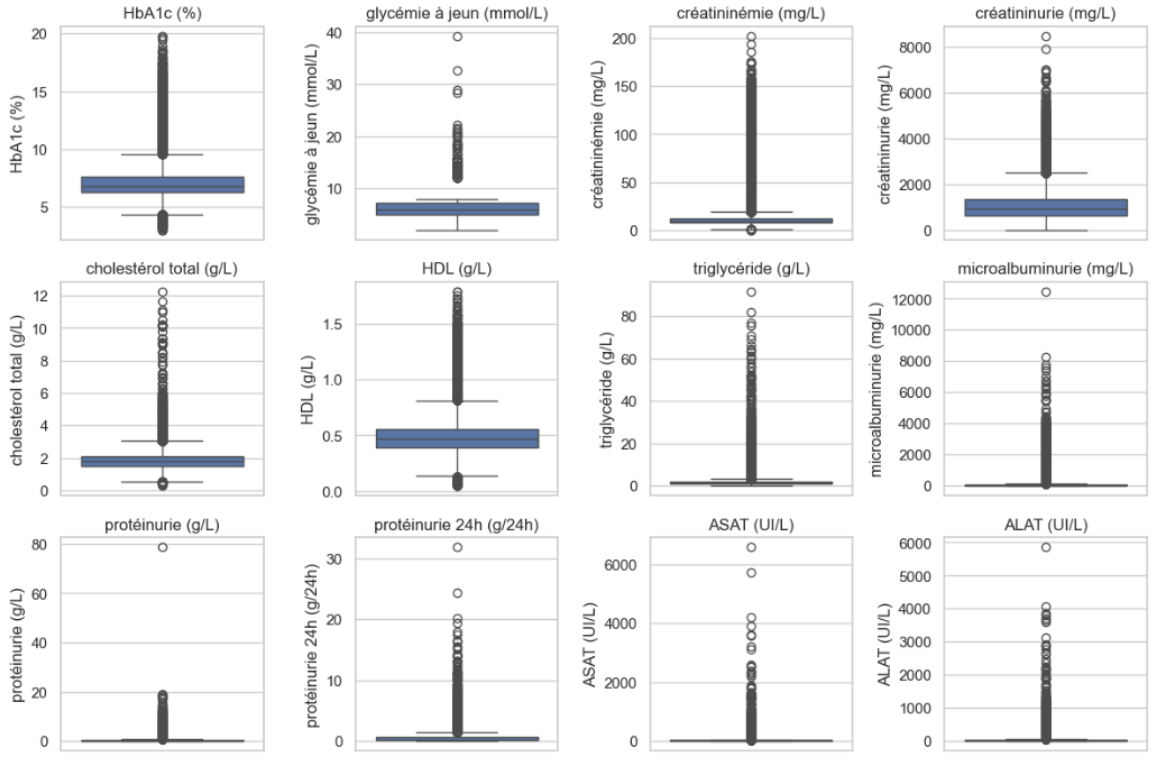
Seules les observances des **taux de microalbuminurie**, de **protéinurie** et de **protéinurie sur 24 heures**, ainsi que les **taux d'ALAT** et **d'ASAT**, montrent une forte dépendance positive respectivement. En effet, plus le **taux d'ALAT** augmente, plus le **taux d'ASAT** a tendance à augmenter. De même, plus le **taux de microalbuminurie** augmente, plus les **taux de protéinurie** et de **protéinurie sur 24 heures** ont tendance à augmenter.

Le **taux de HbA1c** et le **taux de glycémie** ont une corrélation de 0.45, ce qui peut être considéré comme une forte dépendance. Cela signifie que lorsque le **taux de HbA1c** augmente, le **taux de glycémie** a également tendance à augmenter.

1.2.3 La distribution des variables



Le **Taux HbA1c**, **Taux glycémie**, **Taux cholestérol** total et le **Taux HDL** semblent suivre, presque, une loi normale. Cequi veut dire que pour ces obseverances, les valeurs des patiens tournent autour de leurs moyenne.

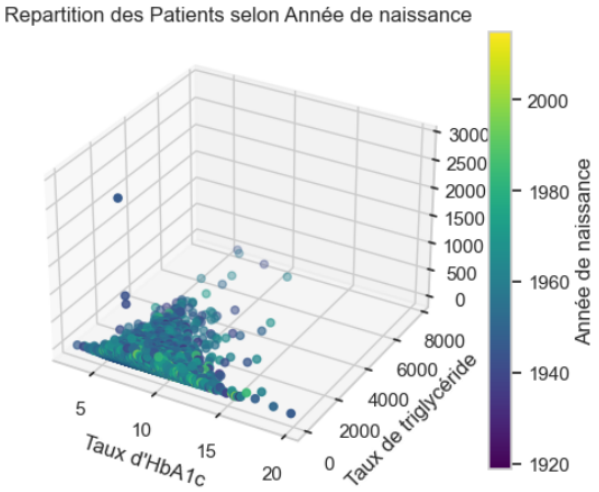


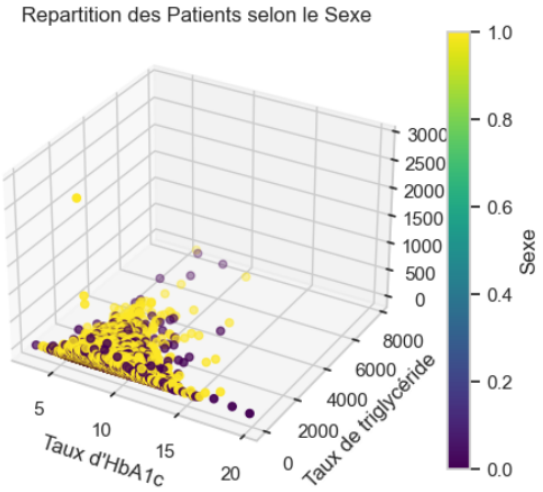
En examinant les box plots des différentes variables, nous remarquons la présence de nombreuses valeurs aberrantes pour chacune d'elles. Ces valeurs aberrantes peuvent avoir un impact significatif sur les analyses statistiques et les modèles de partitonnement que nous aurons à faire si elles ne sont pas traitées correctement. Par conséquent, avant de procéder au partitionnement des patients, il sera crucial de normaliser ces variables.

1.2.4 Répartition des patients selon leurs sexes et leurs années de naissance

Cette analyse nous permet de mieux comprendre quels types de patients sont les plus touchés par le diabète, en examinant les tendances selon le sexe et les périodes de naissance. Avec la corrélation que nous avons détectée entre le taux d'ALAT et le taux d'ASAT, nous pouvons dire que l'un de ces deux taux peut représenter l'information de l'autre dans l'analyse globale des patients. C'est également le cas pour les taux de microalbuminurie, de protéinurie et de protéinurie sur 24 heures, qui sont fortement corrélés.

Ainsi, nous visualisons la répartition des patients dans un espace 3D avec pour axes le taux de HbA1c, le taux d'ALAT et le taux de microalbuminurie.





Nous pouvons remarquer que le diabète touche surtout les **hommes**. On peut remarquer qu'il y'a très très peu des patients nés aux années 2000. Surtout nous avons afffaire à des personnes des années 60 à 80.

1.2.5 Etude des valeurs manquantes

Analyse cruciale pour les variables qui seront utiles pour les méthodes de partitionnement.

```
code patient                     0
date dossier                     0
mois de naissance              131
année de naissance             131
sexe                            47
code postal tronqué             12
HbA1c (%)                   343486
glycémie à jeun (mmol/L)    837668
créatininémie (mg/L)        539088
créatininurie (mg/L)        746385
cholestérol total (g/L)     505776
HDL (g/L)                   511801
triglycéride (g/L)          502548
microalbuminurie (mg/L)     779594
protéinurie (g/L)           807208
RAC (mmol/L)                812109
protéinurie 24h (g/24h)     837188
ASAT (UI/L)                 470781
ALAT (UI/L)                 470751
dtype: int64
```

Sur une base de **844170 observations**, on constate que les variables **glycémie à jeun (mmol/L)**, **créatininurie (mg/L)**, **microalbuminurie (mg/L)**, **protéinurie (g/L)**, **RAC (mmol/L)**, **protéinurie 24h (g/24h)** ont été consultées chez les patients presque exclusivement (très très peu). Nous allons écarter ces variables car la gestion de leurs valeurs manquantes pourrait très probablement biaiser l'étude.

Les obseverances biologique qui seront abordées sont alors les variables **HbA1c (%)**, **ALAT (UI/L)**, **ASAT (UI/L)**, **triglycéride (g/L)**, **cholestérol total (g/L)**, **HDL (g/L)**.

Cependant, nous ne perdons pas beaucoup d'information puisque on a vu sur la partie de l'étude des correlation, par exemple, le taux de HbA1c est corrélé avec le taux de glycémie, alors l'information du taux de glycémie est implicitement contenue dans le taux de HbA1c .

Remarque :

→ Comme mentionné lors de la présentation de la base de données pour l'étude sur le diabète, nous mettons l'accent principalement sur le taux de HbA1c. Par conséquent, nous envisageons de filtrer les patients qui n'ont aucune consultation enregistrée pour la variable **HbA1c (%)**.

→ De même, nous excluons également les patients qui ont moins de 3 consultations, car leur suivi diabétique au fil du temps n'est pas pertinent pour notre analyse.

Important :

Il est essentiel de noter que notre objectif de partitionnement ne concerne pas les observations individuelles, mais plutôt les patients eux-mêmes. Étant donné qu'un patient peut avoir plusieurs consultations, cela se traduit par des entrées répétées pour un même patient dans la base de données. Ainsi, notre approche consiste à créer une base où chaque patient est représenté une seule fois comme une observation.

Dans cette optique, pour chaque patient, nous attribuons à chaque variable la dernière valeur enregistrée lors de la dernière consultation du patient. Cela garantit que chaque patient est représenté de manière cohérente dans notre analyse, en tenant compte de l'évolution de ses paramètres médicaux au fil du temps.

Au passage, nous avons trier les colonnes qui seront utiles au partitionnement des patients, à savoir HbA1c (%) , ALAT (UI/L), ASAT (UI/L) , triglycéride (g/L) , cholestérol total (g/L) et HDL (g/L) .

À l'issue de cette étape, nous disposons d'une base de données où chaque patient est représenté une seule fois. Cette base contient les valeurs manquantes suivantes pour chaque patient :

```
HbA1c (%)                  0
ALAT (UI/L)             1878
ASAT (UI/L)             1877
triglycéride (g/L)       892
cholestérol total (g/L)  914
HDL (g/L)                974
dtype: int64
```

La présence de valeurs manquantes indique que certains patients n'avaient aucune donnée enregistrée pour certaines variables lors de leurs consultations. Ainsi, lorsqu'on attribue la dernière valeur de la dernière consultation, cela peut conduire à des valeurs manquantes.

Pour contourner ce problème, une approche d'imputation pertinente pourrait consister à imputer les valeurs en fonction des groupes de patients avec **diabète contrôlé** et **non contrôlé**. Cela permet de mieux respecter la variabilité entre les patients et pourrait potentiellement améliorer la pertinence des imputations. Nous choisissons d'imputer les valeurs manquantes par la médiane de chaque variable, séparément pour les deux groupes de patients.

```
[27]: Table.isna().sum()
```

```
[27]: HbA1c (%)                0
      ALAT (UI/L)              0
      ASAT (UI/L)              0
      triglycéride (g/L)       0
      cholestérol total (g/L)  0
      HDL (g/L)                0
      Niveau diabète           0
      dtype: int64
```

```
[28]: Table.head()
```

[28]:

	HbA1c (%)	ALAT (UI/L)	ASAT (UI/L)	triglycéride (g/L)	cholestérol total (g/L)	HDL (g/L)	Niveau diabète
0	6.6	27.0	43.0	0.75	1.06	0.48	Diabète contrôlé
1	6.6	24.0	16.0	0.93	1.70	0.64	Diabète contrôlé
2	6.2	22.0	20.0	1.74	1.76	0.35	Diabète contrôlé
3	6.9	23.0	18.0	0.82	1.56	0.44	Diabète contrôlé
4	7.5	22.0	13.0	1.59	1.96	0.28	Diabète non contrôlé

Nous pouvons regarder la description :

```
[29]: Table.describe()
```

[29]:

	HbA1c (%)	ALAT (UI/L)	ASAT (UI/L)	triglycéride (g/L)	cholestérol total (g/L)	HDL (g/L)
count	43440.000000	43440.000000	43440.000000	43440.000000	43440.000000	43440.000000
mean	6.982977	28.709945	27.985083	1.479059	1.705283	0.491683
std	1.226150	28.291066	32.190516	1.017801	0.440439	0.146577
min	3.000000	1.000000	5.000000	0.170000	0.300000	0.050000
25%	6.200000	17.000000	19.000000	0.920000	1.390000	0.390000
50%	6.700000	23.000000	24.000000	1.250000	1.670000	0.470000
75%	7.500000	33.000000	30.000000	1.740000	1.970000	0.570000
max	17.800000	2163.000000	3563.000000	56.160000	10.510000	1.630000

Ainsi, on a constaté 26713 patients à diabète controlé et 16727 patients à diabète non controlé.

1.2.6 Normalisation de données

Cette étape est crucial pour l'application des méthodes de partitionnement. En effet, cette étape est particulièrement importante si des valeurs aberrantes sont présentes. Nous avons constaté dans la partie distribution de données notamment sur les BoxPlot que des valeurs aberantes se presentent, alors ramener toutes les données sur la même échelle permet de garantir que toutes les variables contribuent de manière équitable aux analyses et évite que certaines variables ne dominent autres en raison de leurs echelles plus grandes.

2 Les méthodes de partitionnement

Les méthodes de partitionnement visent à regrouper les patients présentant des similitudes en fonction des observances médicales que nous avons retenues. Nous abordons deux méthodes de partitionnement :

2.1 Classification Ascendante Hiérarchique (CAH)

Le Clustering CAH est une méthode qui crée une hiérarchie de groupe en utilisant une approche agglomérative. Pour appliquer la Classification Ascendante Hiérarchique (CAH), nous parcourons 6 étapes :

2.1.1 Choix du critère de dissimilarité

Dans notre analyse descriptive, nous avons constaté que les variables sélectionnées, celles rétenues pour le partitionnement, n'ont pas de corrélations très significatives entre elles. Bien que ce soient des variables continues, nous devons choisir une mesure de dissimilarité adaptée. La distance euclidienne est la candidate idéale pour les variables continues car elle quantifie la distance linéaire directe entre deux points dans l'espace multidimensionnel.

2.1.2 Choix de la méthode d'agrégation des clusters

Les méthodes d'agrégation disponibles incluent **single** (liaison simple), **complete** (liaison complète), **average** (liaison moyenne), et **ward** (méthode de Ward). Nous utilisons le critère de l'**inertie intra-cluster** pour comparer ces méthodes d'agrégation.

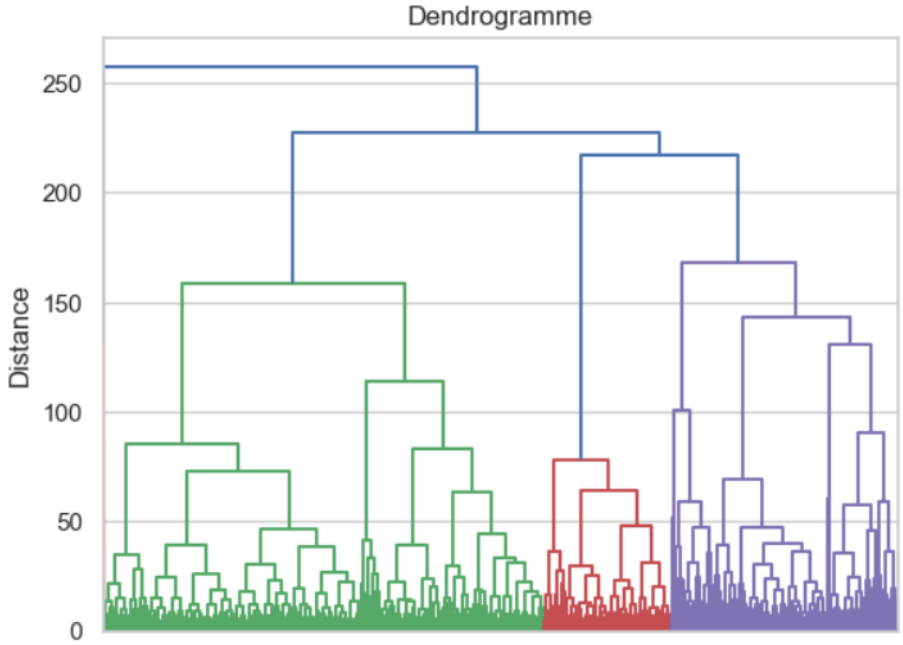
La méthode qui **minimise l'inertie intra-cluster** est généralement préférée car elle conduise à des clusters plus compacts et homogenes.

```
Pour la méthode ward, on a :
{'totI': 6.000000000000002, 'intraI': 5.235711588229425, 'interI': 0.7642884117705764}
Pour la méthode single, on a :
{'totI': 6.000000000000002, 'intraI': 5.591202476303766, 'interI': 0.40879752369623557}
Pour la méthode complete, on a :
{'totI': 6.000000000000002, 'intraI': 5.299481519380927, 'interI': 0.700518480619075}
Pour la méthode average, on a :
{'totI': 6.000000000000002, 'intraI': 5.591202476303766, 'interI': 0.40879752369623557}
```

Sur la base de ces résultats, la **méthode ward** semble être la meilleure option car elle minimise l'inertie intra-cluster tout en maximisant l'inertie inter-cluster, ce qui devrait donner des clusters plus compacts et mieux définis par rapport aux autres méthodes.

2.1.3 Classification Ascendante Hiérarchique (CAH) selon critère de dissimilarité et la méthode d'agrégation des clusters retenus

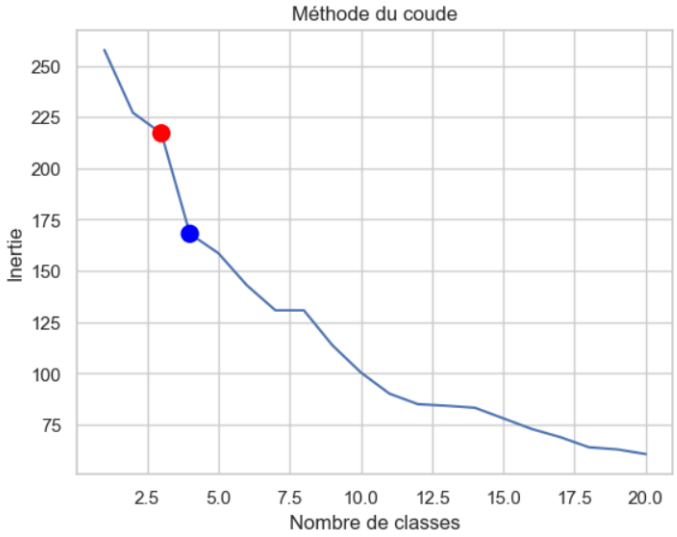
Nous allons donc utiliser la méthode de **Ward** pour construire un modèle de Classification Ascendante Hiérarchique (CAH) avec la **distance eucludienne** comme mésure de dissimilarité.



2.1.4 Sélection d'un nombre de classes adéquat : coude et score de silhouette

Courbe d'inertie en fonction des groupes : Méthode de coude

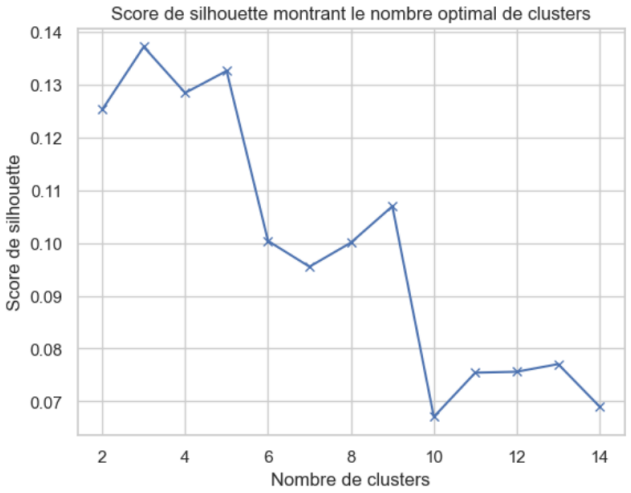
Cette méthode consiste à examiner la courbe d'inertie et à identifier le point où il y a une rupture nette ou un **coude**. Ce point correspond généralement au nombre optimal de classes. L'idée est de sélectionner le nombre de classes juste avant le point où l'inertie cesse de diminuer de manière significative.



Avec ce resultat, on constate une perte importante d'inertie entre $K = 3$ et $K = 4$. Ceci suggere que le nombre optimal de classe qu'on peut retenir est de $K = 3$. Pour affirmer notre décision, on peut regarder le **score de silhouette**.

Score de silhouette en fonction de groupes CAH

Le **score de silhouette** mesure la cohésion et la séparation des groupes. Un **score de silhouette maximal** indique que les groupes sont bien définis et bien séparés. Plus le **score de silhouette** se rapproche de 1 mieux est le partionnement, cequi construit des groupes plus homogenes et compacts.



Le **score de silhouette** est **maximal** pour $k = 3$, ce qui suggère que **3 groupes** offrent la meilleure structure en termes de cohésion et de séparation des données.

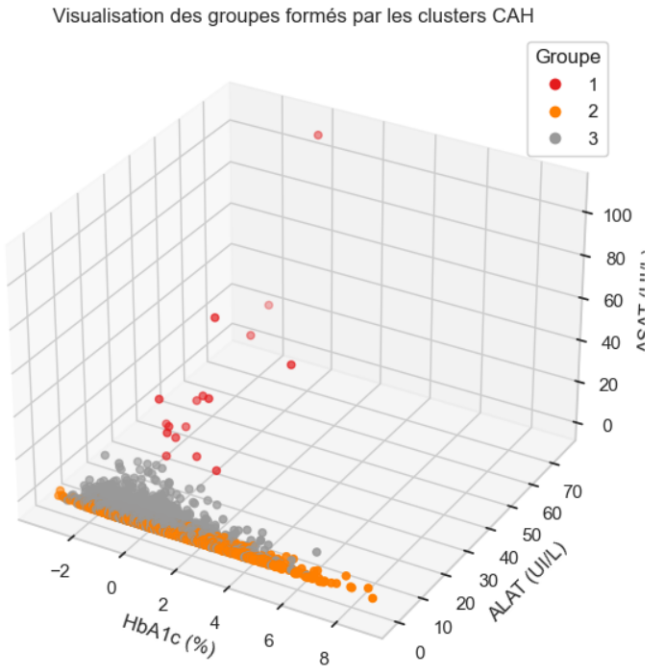
2.1.5 Partitonnement CAH selon le nombre de classe retenue

On applique les **CAH** en trois groupes :

[27]:

	HbA1c (%)	ALAT (UI/L)	ASAT (UI/L)	triglycéride (g/L)	cholestérol total (g/L)	HDL (g/L)	Groupe patient
0	-0.312345	-0.060442	0.466445	-0.716317	-1.465106	-0.079708	2
1	-0.312345	-0.166484	-0.372321	-0.539463	-0.011995	1.011881	3
2	-0.638573	-0.237178	-0.248060	0.256380	0.124234	-0.966624	2
3	-0.067673	-0.201831	-0.310191	-0.647540	-0.329863	-0.352605	2
4	0.421669	-0.237178	-0.465518	0.109002	0.578331	-1.444194	3

Visualisons les groupes formés par les clusters dans un espace 3D avec HbA1c (%), ALAT (UI/L), et ASAT (UI/L).



2.1.6 Description des groupes obtenus par CAH

Ici on regarde la description des groupe selon le **taux de HbA1c**.

```
[45]: moyenne_groupe = Data_base.groupby('Groupe patient')['HbA1c (%)'].describe()
      moyenne_groupe
```

[45]:

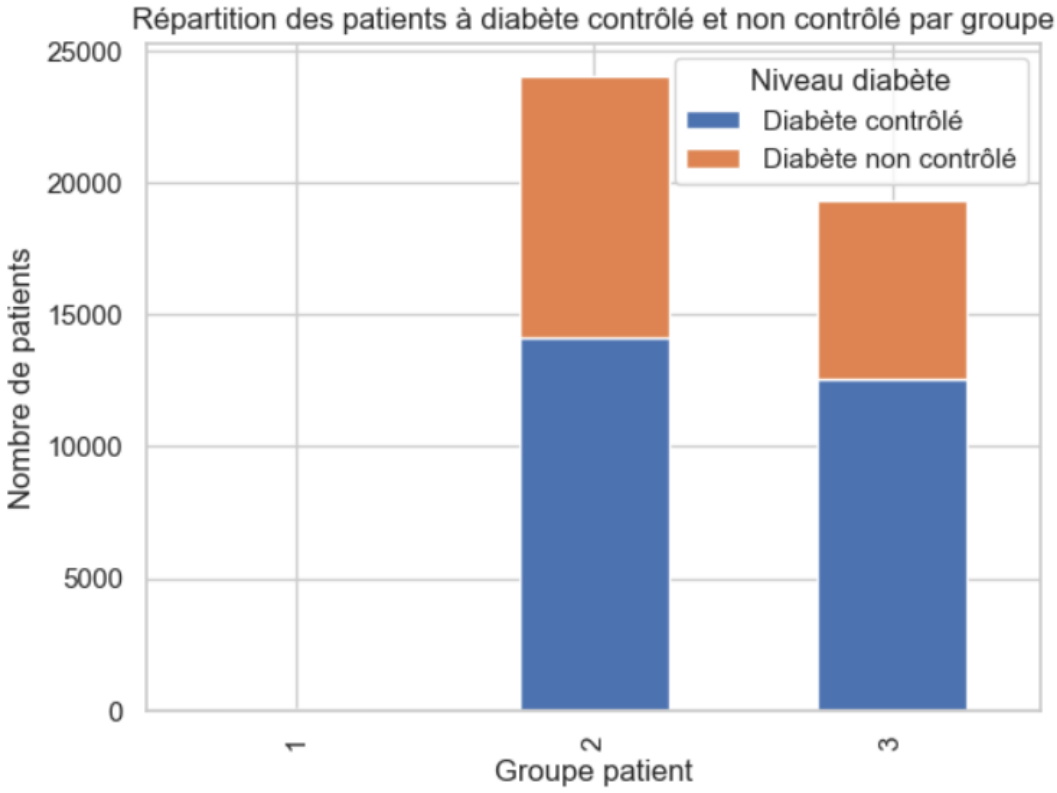
Groupe patient	count	mean	std	min	25%	50%	75%	max
1	17.0	-0.140595	1.056835	-1.698815	-0.736441	-0.475459	0.176998	2.786824
2	24074.0	0.078300	1.052288	-3.248399	-0.638573	-0.149230	0.503226	8.822046
3	19349.0	-0.097297	0.921645	-2.922170	-0.720130	-0.230788	0.340112	6.375334

En regardant la répartition des patients à diabète contrôlé ou non dans les différents groupes :

Groupe patient	Nombre de patients	Diabète contrôlé	Diabète non contrôlé
1	17	12	5
2	24074	14106	9968
3	19349	12595	6754

On peut remarquer que c'est le **groupe 2** qui a beaucoup d'individus d'un nombre de 24074 dont 14106 patients à diabète contrôlé et 9968 patients à diabète non contrôlé.

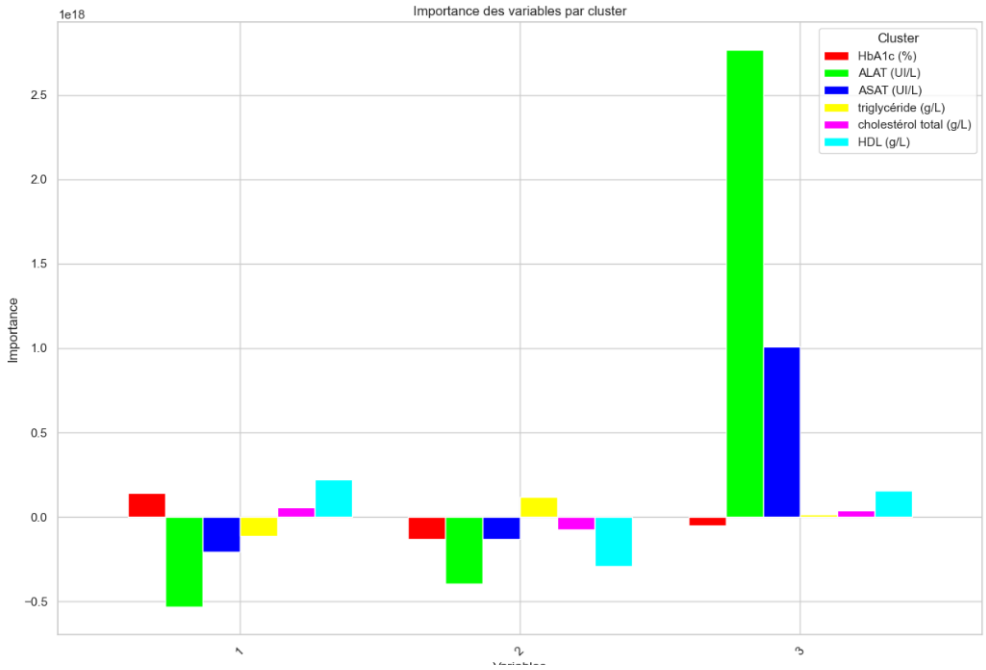
En générale, dans les trois groupes, il y'a autant des patients à diabète contrôlé par rapport aux patients à diabète non contrôlé. On peut le voir dans le graphique ci-dessous :



2.1.7 Contribution des variables pour chaque groupe

Chaque barre dans le graphique en barres représente une variable explicative, et sa hauteur indique l'importance de cette variable pour le cluster correspondant. Pour ce faire on calcule d'abord les

moyennes des variables pour chaque groupe, puis compare ces moyennes à la moyenne globale de chaque variable. L'importance relative de chaque variable est ensuite calculée.



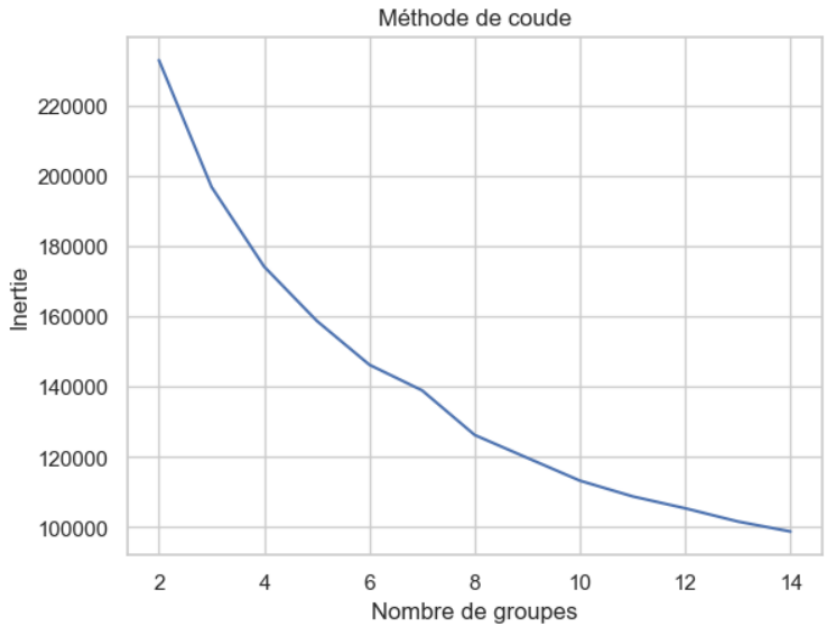
Dans l'ensemble, la variable **taux d'ALAT** est la plus contributive pour le partitionnement surtout dans le **groupe 3**.

2.2 Clustering k-moyennes (K-means)

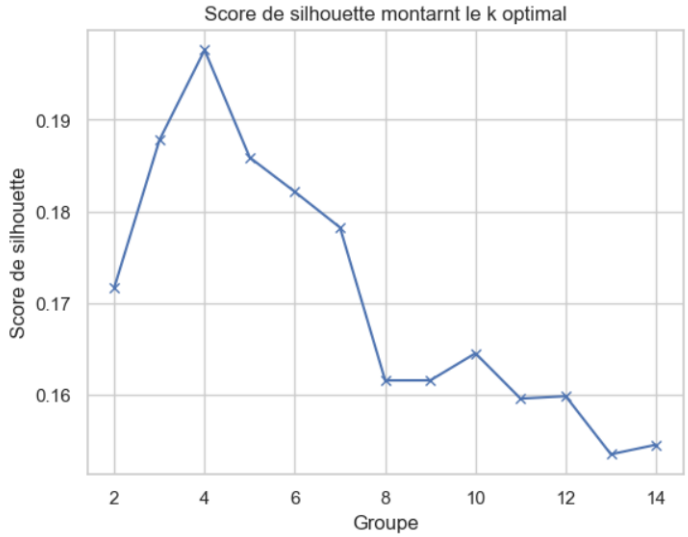
Le **Clustering k-moyennes** regroupe les observations en k groupes en minimisant la variance au sein de chaque groupe. Pour partitionner des groupes les homogenes, héterogene entre groupe et compact, tout comme la méthode de CAH, il y'a des étapes à tenir compte :

2.2.1 Déterminer le nombre de K groupe optimal : coude et score de silhouette

Nous appliquons la méthode de **coude** et le **score de silhouette**.



On constate une perte importante d'**inertie** entre $K = 2$ et $K = 4$. Ceci suggere que le nombre optimal de classe qu'on peut retenir est de $K = 4$ ou $k = 3$. Pour clarifié notre décision, on peut regarder le **score de silhouette**.



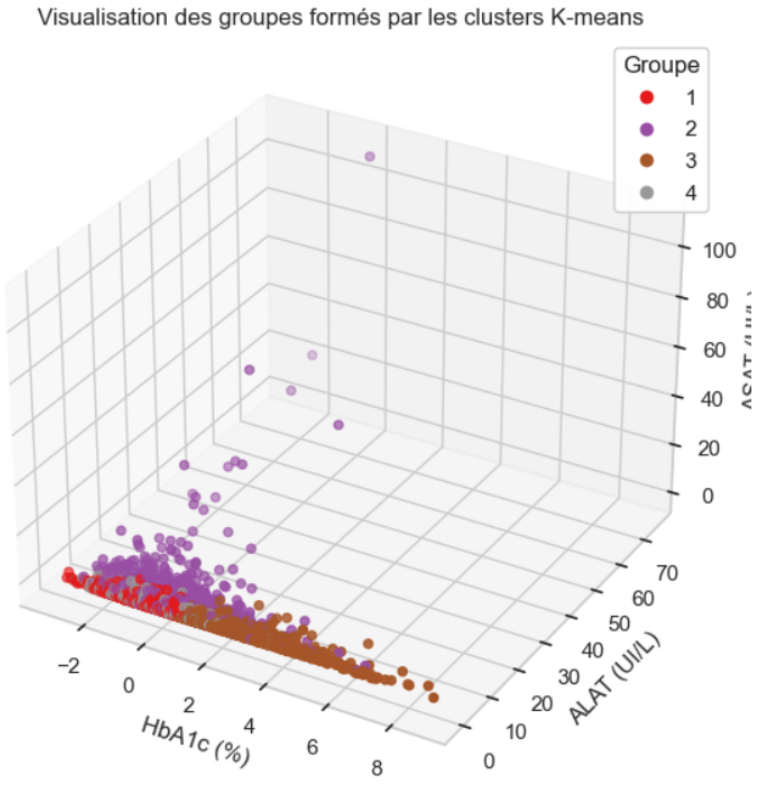
Le **score de silhouette** est maximal pour $k = 4$, ce qui suggère que 4 groupes offrent la meilleure structure en termes de cohésion et de séparation des données.

2.2.2 Partitionnement de k-means avec le nombre de classe obtenu

Finalement, on a appliqué le **K-means** en 4 goupes.

[83]:

	HbA1c (%)	ALAT (UI/L)	ASAT (UI/L)	triglycéride (g/L)	cholestérol total (g/L)	HDL (g/L)	Groupe_Kmeans
0	-0.312345	-0.060442	0.466445	-0.716317	-1.465106	-0.079708	1
1	-0.312345	-0.166484	-0.372321	-0.539463	-0.011995	1.011881	4
2	-0.638573	-0.237178	-0.248060	0.256380	0.124234	-0.966624	1
3	-0.067673	-0.201831	-0.310191	-0.647540	-0.329863	-0.352605	1
4	0.421669	-0.237178	-0.465518	0.109002	0.578331	-1.444194	1



Visualisation des groupes formés par les clusters K-means dans un espace 3D avec HbA1c (%), ALAT (UI/L), et ASAT (UI/L).

Au passage, nous avons calculer les inerties inter et intra clusters :

```
Inerties pour K-means:
{'totI': 5.000000000000002, 'intraI': 3.672172896509314, 'interI': 1.3278271034906877}
```

2.2.3 Description des Groupe de K-means

Ici on regarde la description des groupe selon le **taux de HbA1c**.

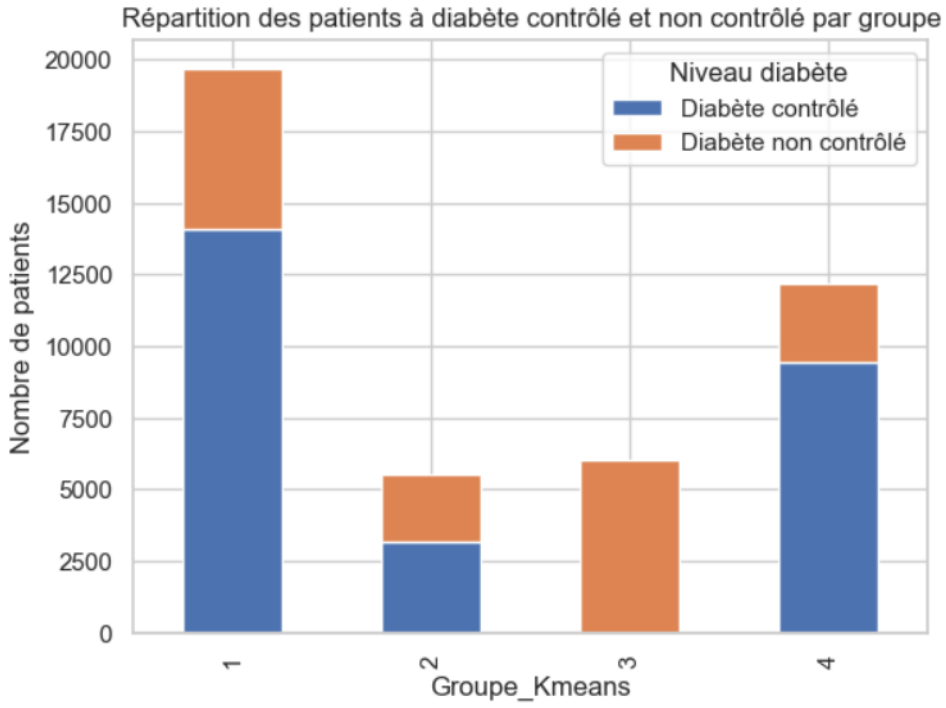
[76]:

Groupe_Kmeans	count	mean	std	min	25%	50%	75%	max
1	19714.0	-0.295193	0.564114	-3.248399	-0.720130	-0.312345	0.095441	1.237240
2	5522.0	0.015191	0.879808	-2.595942	-0.557016	-0.067673	0.421669	6.538448
3	6038.0	1.743629	0.997517	0.421669	1.074125	1.481911	2.134367	8.822046
4	12166.0	-0.393924	0.636097	-2.922170	-0.801687	-0.475459	0.013884	3.683951

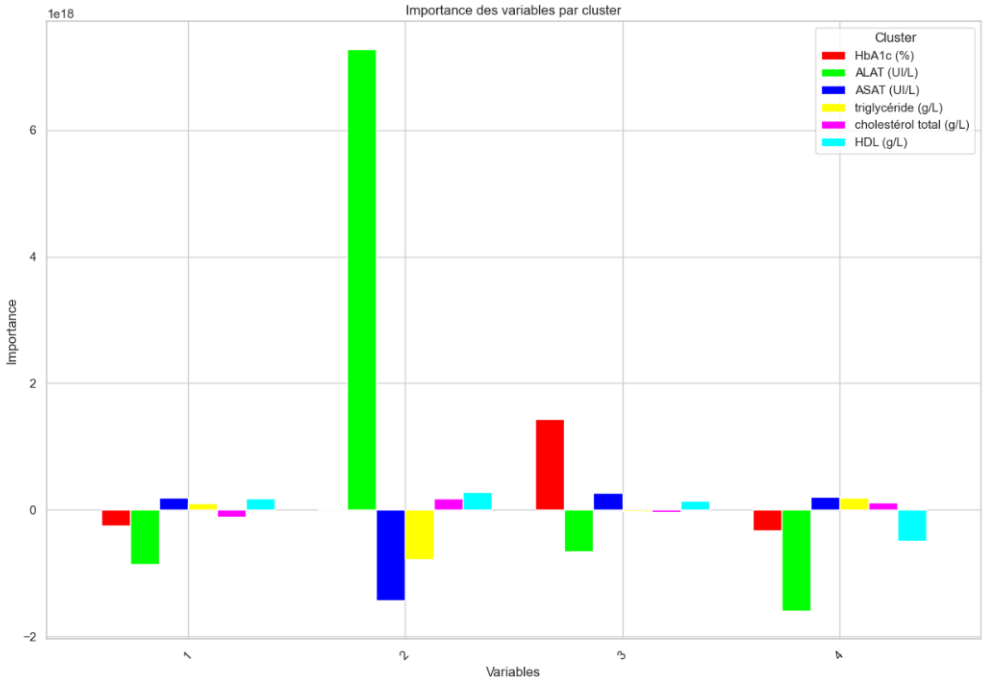
En regardant la repartition des patients à diabète controlé ou non dans les différents groupes.

Groupe_Kmeans	Niveau diabète	Nombre de patients	Diabète contrôlé	Diabète non contrôlé
1		19714	14096	5618
2		5522	3188	2334
3		6038	0	6038
4		12166	9429	2737

En générale, dans les quatres groupes, il y'a autant des patients à diabète controlé par rapport aux patients à diabète non controlé. Cependant, on peut regarder la 3ème groupe composé de 6038 patients qui sont tous des patients à diabète non controlé. On peut le voir dans le graphique ci-dessous :



2.2.4 Contribution des variables pour chaque groupe



On remarque que dans le 1ème, 2ème et 4ème goupes le **Taux d'ALAT** a quasiment été le plus contributif. Quant au **Taux de HbA1c**, il a été plus contibutif dans le 3ème groupe.

3 Conclusion

Les méthodes de partitionnement sont des outils puissants pour regrouper des patients en fonction de leurs observances médicales. Dans cette analyse des patients diabétiques, on constate que le diabète touche principalement les hommes, surtout les personnes nées aux 60 à 80 ans.

Pour former des groupes de patients similaires, nous avons appliqué deux méthodes principales : la **CAH** et le **Clustering K-means**. Voici les principales conclusions :

Classification Ascendante Hiérarchique (CAH) :
Trois groupes (K=3) ont offert la meilleure structure en termes de cohésion et de séparation des données. En appliquant la CAH avec trois groupes, nous avons observé une répartition équilibrée entre les patients avec un **diabète contrôlé** et **non contrôlé** dans chaque groupe. Le groupe 2 s'est distingué par le plus grand nombre de patients, tandis que le groupe 1 s'est distingué par le plus faible nombre de patients. Quant à la contribution des variables, le **taux d'ALAT** a été identifié comme la variable la plus contributive pour le partitionnement, surtout dans le groupe 3..

Clustering K-means :
Quatre groupes (K=4) ont offert la meilleure structure en termes de cohésion et de séparation des données. En appliquant K-means avec quatre groupes, nous avons observé une répartition distincte des patients. Notamment, le groupe 3 était composé uniquement de patients avec un **diabète non contrôlé**. Quant à la contribution des variables, le **taux d'ALAT** a été la variable la plus contributive dans les groupes 1, 2 et 4, tandis que le **taux de HbA1c** était plus contributif dans le groupe 3.

En conclusion, les deux méthodes ont leurs avantages spécifiques. La CAH, avec la méthode de Ward, offre des clusters bien définis et compacts, tandis que K-means peut identifier des sous-groupes distincts avec des caractéristiques médicales spécifiques. En examinant la visualisation des clusters formés par les deux méthodes dans l'espace 3D formé par **HbA1c (%)**, **ALAT (UI/L)** et **ASAT (UI/L)**, la **CAH** semble fournir le meilleur partitionnement.

4 Références

Nous nous sommes inspiré par des cours, tutoriel et article, notament :

→ Cours de Data Mining et Analyse de données dispensés Master 1 MIASHS, Univ Lille 1.

→ Tutoriel du Youtubeuse **LeCoinStat**, via le lien : Le Guide Complet pour Maîtriser K-means en Python

→ Tutoriel du Youtubeuse **LeCoinStat**, via le lien : Classification Ascendante Hiérarchique (CAH) en Python - Jour 83

→ Un article publié en Novembre 2023, via le lien : Institut Amelis - Comprendre le diabète : types, causes, symptômes et traitements

→ L'HBA1C OU HÉMOGLOBINE GLYQUÉE, via le lien : Fédération Française des Diabètes.

→ DIABÈTE : L'IMPORTANCE DE L'HÉMOGLOBINE GLYQUÉE, via le lien : Gifar.