République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Blida -1-

Faculté des Sciences

Département d'Informatique

Mémoire de fin d'étude présenté en vue d'obtention du diplôme

MASTER EN INFORMATIQUE

Option : Systèmes Informatiques & Réseaux

Thème

Classification hiérarchique ascendante non supervisée des commandes de la société CASBAH

Présenté par : TABOUCHE MOHAMED Amine

Soutenus le 10/ 2021, Devant les jurys :

Mr. KAMECHE A.	**(Président)**
Mme. Lahiani N.	**(Examinatrice)**
Mr. DOUGA Yassine.	**(Promoteur)**

Promotion: 2020/2021

Remerciement

Avant tout nous formulons notre gratitude à Allah le tout puissant de nous avoir donné la force d'achever ce travail,

Nous tenons aussi à remercier : Mr. Douga notre promoteur, pour ses conseils,

ses orientations, et sa disponibilité,

qui nous ont permis de réaliser ce travail

dans les meilleures conditions,

Ainsi les membres du jury, pour avoir fait l'insigne honneur d'accepter de lire et juger ce travail,

Un très grand merci à tous les enseignants de notre département, qui ont assuré notre enseignement pendant tout notre parcours académique.

Sans oublier de remercier nos parents

pour leur contribution, leur soutien et leur patience, nos proches,

nos amis, nos collègues

et toutes personnes qui nous ont aidées par leur soutien permanent

de près ou de loin de nos études.

ملخص

تستند إدارة طلبات الشركة ومعالجتها إلى قاعدة بيانات ، لذلك لن تكون هناك مشكلة في عدد الطلبات حيث يستخدم التصنيف لاستخراج المعلومات من البيانات الضخمة باستخدام تعلم الكمبيوتر. لقد ذكرنا طريقتين رئيسيتين للتصنيف الخاضع للإشراف وغير الخاضع للإشراف. إحدى الطريقتين الأساسيتين للتصنيف غير الخاضع للرقابة هي التصنيف الهرمي ، والذي يتكون من طريقتين رئيسيتين ، وهما التصنيف الهرمي التصاعدي والتنازلي. تختلف طرق التقسيم الطبقي هذه الأخيرة عن بعضها البعض في اختيار معايير التشابه.وكذلك درسنا طريقة تصنيف مثل "طريقة وارد" ، وطبقناها لحل مشاكل في قاعدة بيانات كبيرة جدًا ويصعب إدارتها.

الكلمات المفتاحية:

قاعدة البيانات ، التصنيف ، طريقة وارد.

Résumé

La gestion et le traitement internes des commandes de l'entreprise utilisent tous la base de données, il n'y aura donc aucun problème avec le nombre de commandes. La classification est utilisée pour extraire des informations des big data à l'aide de l'apprentissage informatique. Les deux principales méthodes sont la classification supervisée et non supervisée. L'une des deux méthodes de base de la classification non supervisée est la classification hiérarchique, qui comporte deux méthodes principales, à savoir la classification hiérarchique ascendante et descendante. Ces méthodes de stratification diffèrent les unes des autres dans la sélection des critères de similarité. L'une des méthodes de calcul de la similarité est la méthode "Ward".

Mots clés :

Base des données, classification, méthode de Ward.

Abstract

The management and the processing of the orders within the companies is based on the databases and not to have problems compared to the numerous orders. A classification is used to extract information from big data using computer learning. We cite the two main methods of supervised and unsupervised classification and one of the two basic methods of unsupervised classification are hierarchical classification which includes two main methods namely ascending and descending hierarchical classification. These hierarchical methods differ from each other in the choice of the similarity criterion. In this work, we will study a classification method such as "Ward's method" and apply it to solve their problems in a very large and difficult to manage the database.

Keywords:

Database, classification, ward's method.

Table des matières

INTRODUCTION GÉNÉRALE 1
PARTIE I : ETAT DE L'ART 3
CHAPITRE I : CLASSIFICATION 4
1 Introduction 5
2 La classification 5
2.1 Historique 5
2.2 Définition 5
2.3 L'objectif de la classification 5
2.4 Formule mathématique 6
2.5 Les méthodes de classification 6
2.5.1 La classification supervisée 7
2.5.2 La classification non supervisée 10
4.Conclusion 19
CHAPITRE II : APPROCHES D'AGRÉGATIONS 20
1 Introduction 21
2 Les différentes approches ou stratégie d'agrégation : 21
3 Comparaison générale et discussion 25
3 Conclusion : 26
PARTIE II : CONCEPTION 27
1 Introduction 28
2 L'algorithme Ward-Linkage 28
2.1 Introduction 28
2.2 Définition 28
2.3 Principe 30
2.4 Organigramme d'algorithme de clustering à base de Ward 31
2.5 Fonction 32
3 Conclusion 33
PART III : IMPLÉMENTATION ET ÉVALUATION 34
1 Introduction 35

2 Matériel informatique (Hardware) 35
3 Logiciel (Software) 35
4 Les Bibliothèque 36
5 Base de données (Dataset) 36
6 Implémentation : 37
7 La classification de base de données 38
8 Code source 39
9 Conclusion 42
CONCLUSION GÉNÉRALE 43
BIBLIOGRAPHIE 45

Liste des Figures

Figure 1: Les méthodes de classification.[5]..... 7
Figure 2: Classification selon la méthode hiérarchique..... 12
Figure 3 : La première étape de CAH..... 14
Figure 4 : La deuxième étape de CAH 15
Figure 5 : La troisième étape de CAH..... 16
Figure 6:Différents étapes jusqu'aux résultats final [15]..... 17
Figure 7: Dendrogramme représente résultat final. [15]..... 18
Figure 8: Schémas de la classification descendante. 19
Figure 9: Schéma de la classification « Single-Linkage ». 22
Figure 10: Schéma de la classification « Complete-Linkage ». 23
Figure 11: Schéma de la classification « Average-Linkage ». 23
Figure 12: Schéma de la classification «Centroid -Linkage ». 24
Figure 13: Schéma de la classification «Ward-Linkage »[15]. 24
Figure 14: La différence entre les quatre méthodes dans le résultat final (dendrogramme de chaque algorithme)[15]. 26
Figure 15: Méthode de Ward [15]. 30
Figure 16: Organigramme de l'algorithme de Ward. 32
Figure 17: Représente les importantes tables dans notre système. 37
Figure 18: Représente la commande SQL pour créer la vue..... 37
Figure 19: La méthode Ward-linkage [15]. 38
Figure 20 : Importation des bibliothèques 39
Figure 21: Importation de base des données 39
Figure 22: Vérification le nombre d'attributs 39
Figure 23: Commande de choisir des colonnes..... 39
Figure 24:Commande pour applique la méthode de Ward et affiche le dendrogramme 39
Figure 25: Dendrogramme représente le résultat de méthode de Ward 40
Figure 26 : Commande pour affiche le résultat des points de données 40
Figure 27:Tracer les clusters. 41
Figure 28: Résultat finale représente par des points de données. 41

Liste des tableaux

Tableau 1: Comparaison entre les différents algorithmes de CAH. 25-26

Liste des abréviations :

CAH : Classification Ascendant Hiérarchies.

CDH : Classification Descendante Hiérarchies.

SVM : Support Vecteur Machine.

SARL : Société A Responsabilité limitée.

SQL: Structured Query Language.

SGBD: Système de Gestion de Base de Données.

K-PPV: K-Plus Proches Voisins.

K-NN: K-Nearest- Neighbors.

CHAID: Chi-squared Automatic Interaction Detector.

CART: Classification And Regression Tree.

TD : Table des Distances.

INTRODUCTION GENERALE

Introduction

Aujourd'hui, les entreprises se concentrent sur des applications Web pour vendre ou fournir des services. Ces applications utilisent des basées de données pour stocker les informations. Avec la croissance rapide du volume des données, il est de plus en plus difficile pour les administrateurs de gérer et traiter ces bases de données.

SARL CASBAH a été créée en 1998. Elle a commencé par la fabrication du vinaigre et est devenu un leader du marché avec 30% d'actions en 2001. Son capital social a connu des augmentations successives, mais dans un autre part les données à gérer et à traiter ont dépassé la capacité de stockage et de traitement classique des données. Et c'est pour ça la thématique de recherche m'a été proposée afin de trouver une solution pour la gestion des données. [1]

La société SARL CASBAH est l'une des entreprises qui utilise des sites web dans son marketing.

Problématique

Comme cite par avant la société utilise une application e-commerce qui permet de passer des commandes tous les jours, qui fournit beaucoup de données importantes, ce qui est devenu un problème pour traiter dans l'entreprise.

Objectif

Les objectifs visés à travers ce travail sont les suivants :

1. Étude bibliographique sur les différentes méthodes de la classification.
2. Étude bibliographique sur la classification ascendante hiérarchique.
3. Étude comparative entre différentes approches d'agrégation.
4. Réalisation et validation de l'approche proposée.

Organisation du mémoire

La structure de ce document est organisée comme suit :

La première partie est compose en deux chapitre :

Le premier chapitre est un chapitre qu'il représente une vue globale sur la classification et ses différentes méthodes.

Le deuxième chapitre est consacré à la classification ascendante hiérarchique, pour avoir une idée sur les différentes solutions qui existes et pour définir les approches de calcul la distance entre les clusters, et se termine par une étude comparative de ces travaux.

La deuxième partie, ici nous allons plus détails sur la méthode de Ward et définir l'architecture de notre système, avec la spécification des différents besoins du système. Nous allons spécifier et détailler les processus nécessaires de la classification ascendante hiérarchique base sur méthode de Ward par un organigramme.

La troisième partie, nous présenterons les différents outils utilisés pour construire notre système. Nous avons testé la solution du problème de classification ascendante hiérarchique base sur la méthode de Ward.

PARTIE I : ETAT DE L'ART

CHAPITRE I : CLASSIFICATION

1 Introduction

Ce chapitre commence par des informations générales sur les bases de données et leurs différents types, puis passe à la classification des bases de données et de leurs différentes méthodes.

2 La classification

2.1 Historique

En 1813 Augustin Pyramus de Candolle a utilisé pour désigner la science des lois de la classification des formes vivantes selon les critères de regroupement : taille, forme des feuilles, racines, etc. sous le nom Taxinomie (grec : ordre, arrangement et loi).

Linné (en sciences naturelles), et Koppen (classification des climats) en 1911.

Et pour la première fois en 1939 l'utilisation de termes classification avec ces différents algorithmes par Tryon.

Robert R. Sokal et Peter H.A. Sneath présente en 1963 des méthodes quantitatives appliquées à la taxinomie [2].

2.2 Définition

Classifier c'est regrouper entre eux des objets similaires selon certain critères par les diverses techniques de classification visent toutes à répartir n individus caractérisés par p variables$X_1, X_2,\ldots, X_p$ en un certain nombre m de sous-groupes aussi homogènes que possible. Selon la méthode de classification qui peut être directe en un nombre fixé de classes ou sous la forme d'une hiérarchie à plusieurs niveaux d'agrégation. Le modèle général s'appuie sur la distance entre un individu et un autre. Plus cette distance est réduite, plus les deux entités sont proches et la classification se fait sur cette base quel que soit la méthode utilisée, ce critère de regroupement ou la nature de distance utilisée. [2]

2.3 L'objectif de la classification

La classification a pris aujourd'hui une place importante en analyse des données exploratoire et décisionnelle, l'objectif exploratoire vise à découvrir une partition hypothétique dans un ensemble d'objets. Dans l'analyse décisionnelle, on cherche généralement à affecter tout nouvel objet à des groupes préalablement définis.

La classification a pour but plus simple est répartir l'échantillon en groupes d'observation homogènes, chaque groupe étant bien différencié des autres.

On veut en général obtenir des sections à l'intérieur des groupes principaux, puis des subdivisions plus petites de ces sections, et ainsi de suite. En bref, on désire avoir une hiérarchie de plus en plus fine, sur l'ensemble d'observations initiales [3].

2.4 Formule mathématique

En terme mathématique, un problème de classification comporte les ingrédients suivants :

- Une population de N individu I^i(i variant de 1 à N)
- P variables descriptives X_d^iqui permettent de décrire les individus ; elles sont aussi appelées plus simplement descripteurs (d variant de 1 à P)
- C classes $C_k$dans lesquelles on cherche à ranger les individus (k variant de1 à C)

Résoudre un problème de classification, c'est trouver une application de l'ensemble des objets à classer, décrits par les variables descriptives choisies, dans l'ensemble des classes [4].
L'algorithme ou la procédure qui réalise cette application est appelé classifier.

2.5 Les méthodes de classification

Il existe un grand nombre de méthodes et surtout beaucoup de variantes. Il est d'objectif de les différencier grossièrement soit par leur structure de classification, soit par le type de représentation des classes.

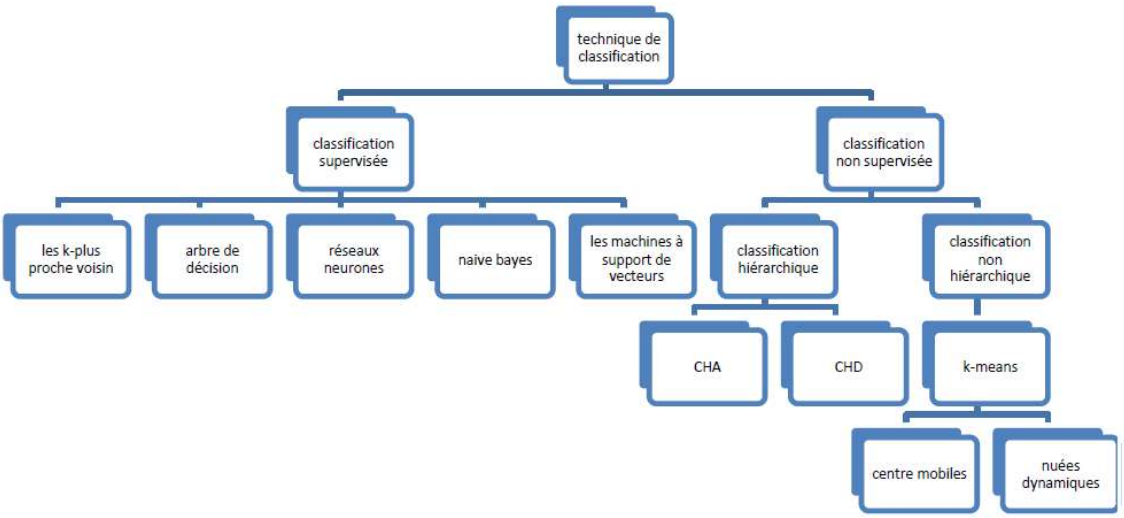

Figure 1: Les méthodes de classification. [5][6] [11]

Ainsi, nous distinguer selon les critères deux types :

2.5.1 La classification supervisée

Dans le contexte supervisé on dispose déjà d'exemples dont la classe est connue et étiquetée. Les données sont donc associées à des labels des classes notées

Q = { q_1 , q_2,….., q_n }. L'objectif est alors d'apprendre à l'aide d'un modèle d'apprentissage des règles qui permettent de prédire la classe des nouvelles observations ce qui revient à déterminer une fonction Cl qui à partir des descripteurs (D) de l'objet associe une classe qi et de pouvoir aussi affecter toute nouvelle observation à une classe parmi les classes disponibles. Ceci revient à la fin à trouver une fonction qu'on note Ys qui associe chaque élément de X un élément de Q. On construit alors un modèle en vue de classer les nouvelles données. Parmi les méthodes supervisées on cite : les k-plus proches voisins, les arbres de décision, les réseaux de neurones, les machines à support de vecteurs (SVM) et les classificateurs de Bayes.

Quel que soit le type de la classification, on est confronté à des problèmes. Dans le cas supervisé, un problème important peut-être le manque de données pour réaliser l'apprentissage ou la disponibilité de données inadéquates par exemple incertaines et imprécises ce qui empêche la construction d'un modèle correct [5].

❖ **Les K-plus proches voisins**

La méthode des k-plus proches voisins (noté K-PPV ou K-NN pour K-Nearest- Neighbors en anglais) consiste à déterminer pour chaque nouvel individu que l'on veut classer, la liste des k plus proches voisins parmi les individus déjà classés. L'individu est affecté à la classe qui contient le plus d'individus parmi ces k plus proches voisins. Cette méthode nécessite de choisir une distance (la plus classique est la distance Euclidienne), et donc le nombre k de voisins à prendre en compte.

Cette méthode supervisée et non-paramétrique est souvent performante. De plus, son apprentissage est assez simple [6].

❖ **Arbre de décision**

Les arbres de décision sont les plus populaires des méthodes d'apprentissage. L'apprentissage se fait par partitionnement récursif selon des règles sur les variables explicatives suivant les critères de partitionnement et les données, on dispose de différentes méthodes, dont CART, CHAID…. Ces méthodes peuvent s'appliquer à une variable. Deux types d'arbres de décisions sont ainsi définis [7] :

- ✔ **Arbre de classification** la variable expliquée est de type nominal. A chaque étape du partitionnement, on cherche à réduire l'impureté totale des deux nœuds fils par rapport au nœud père.
- ✔ **Arbre de régression** la variable expliquée est de type numérique et il s'agit de prédire une valeur la plus proche possible de la vraie valeur.

Construire un tel arbre consiste à définir un nœud, chèque nœud permettant de faire une partition des objets en 2 groupes sur la base d'une des variables explicatives. Il convient donc :

- Définir un critère permettant de sélectionner le meilleur nœud possible à une étape donnée.
- De définir quand s'arrête le découpage, en définissant un nœud terminal (feuille).
- D'attribuer au nœud terminal la classe ou la valeur la plus probable.
- D'élaguer l'arbre quand le nombre de nœuds devient trop important en sélectionnant un sous arbre optimal à partir de l'arbre maximal.
- Valider l'arbre à partir d'une validation croisée ou d'autres techniques.

❖ **Réseaux neurones**

Les réseaux de neurones sont des approximateurs universels parcimonieux ; ils peuvent donc être utilisés pour modéliser ou commander tout processus, statique ou dynamique, non linéaire : en raison de leur parcimonie, ils sont avantageux par rapport aux autres approximateurs et notamment au flou - dès que le processus à modéliser ou à commander possède plus de deux ou trois entrées. Néanmoins, comme toute autre technique, les réseaux de neurones sont soumis à des contraintes : étant des outils statistiques, ils traitent uniquement de données *numériques,* dont le nombre et la représentativité doivent être convenables même si, leur parcimonie leur permet d'utiliser moins de données que d'autres méthodes statistiques. S'il est possible de tirer profit, pour la conception du réseau, des connaissances, même imprécises, que l'on peut avoir sur le processus, il faut qu'elles soient sous forme mathématique : les réseaux de neurones ne permettent pas de traiter aisément des données linguistiques. [8]

- Naïve bayes

  Nommées d'après le théorème de Bayes, ces méthodes sont qualifiées de "naïve" ou "simple" car elles supposent l'indépendance des variables. L'idée est d'utiliser des conditions de probabilité observées dans les données.

  On calcule la probabilité de chaque classe. [9]

- Les machines à support de vecteurs

  Cette technique - initiée par Vapnik - tente de séparer linéairement les exemples positifs des exemples négatifs dans l'ensemble des exemples. Chaque exemple doit être représenté par un vecteur de dimension n. La méthode cherche alors l'hyperplan qui sépare les exemples positifs des exemples négatifs, en garantissant que la marge entre le plus proche des positifs et des négatifs soit maximale. Intuitivement, cela garantit un bon niveau de généralisation car de nouveaux exemples pourraient ne pas être trop similaires à ceux utilisés pour trouver l'hyperplan mais être tout de même situés franchement d'un côté ou l'autre de la frontière. L'efficacité des SVM est supérieure à celle de toutes les autres méthodes sur la classification de textes. Son efficacité est aussi très bonne pour la reconnaissance de formes. Un autre intérêt est la sélection de Vecteurs Supports qui représentent les vecteurs discriminants grâce auxquels est déterminé l'hyperplan.

Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas. Cela en fait une méthode très rapide. [10]

2.5.2 La classification non supervisée

Cette classification est aussi appelée "classification automatique", "clustering" ou encore "regroupement". Dans ce type de classification, on est amené à identifier les populations d'un ensemble de données. On suppose qu'on dispose d'un ensemble d'objets que l'on note par X = $\{x_1, x_2, \ldots\ldots, x_n\}$ caractérisé par un ensemble de descripteurs D, l'objectif du clustering est de trouver les groupes auxquels appartiennent chaque objet x qu'on note par C = $\{C_1, C_2, \quad C_n\}$ Ce qui revient à déterminer une fonction notée Y_s qui associe à chaque élément de X un ou plusieurs éléments de C. Il faut pouvoir affecter une nouvelle observation à une classe. Les disponibles ne sont pas initialement identifiées comme appartenant à telle ou telle population. L'absence d'étiquette de classe est un lourd handicap qui n'est que très partiellement surmontable. Seule l'analyse de la répartition spatiale des observations peut permettre de "deviner" où sont les véritables classes.

Parmi les méthodes non supervisées les plus utilisées, citons deux types d'approches : les centres mobiles (k-means) et la classification hiérarchique. [6]

A. Classification non hiérarchique

Classification non hiérarchique ou partitionnement, aboutissant à la décomposition de l'ensemble de tous les individus en m ensemble disjoints ou classes d'équivalence ; le nombre m de classes est fixé.

Le résultat obtenu est alors une partition de l'ensemble des individus, un ensemble de parties, ou classes de l'ensemble I des individus telles que :

-Toute classe est non vide.

-Deux classes distinctes sont disjointes.

-Tout individu appartient à une classe.

Cet algorithme porte le nom d ' "agrégation autour de centres variables". Une version légèrement différente, connue sous le nom de "nuées dynamiques" consiste à représenter

chaque groupe non pas par son centre, mais par un ensemble de points (noyau) choisis aléatoirement à l'intérieur de chaque groupe. On calcule alors une distance "moyenne" entre chaque observation et ces noyaux et l'on procède à l'affectation. [11]

Méthode de K-means

C'est une méthode dont le but est de diviser des observations en k partitions dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche.
Nous citons deux méthodes connues sur le principe de k-means sont :

→ Méthodes de centres mobiles ;

→ Méthodes des nuées dynamiques.

- Méthode de centres mobiles

Cette méthode consiste à construire une partition en k classes en sélectionnant k individus commence, des classes tirées au hasard de l'ensemble d'individus. Après cette sélection, on affecte chaque individu au centre le plus proche en créant k classes, les centres des classes seront remplacés par les centres de gravité et de nouveaux classes seront créés par le même principe.
Généralement la partition obtenue est localement optimale car elle dépend du choix initial des centres. Pour cela les résultats entre deux exécutions de l'algorithme sont significativement variés.

- Méthode de nuées dynamiques

Dans ce cas, le problème posé est la recherche d'une partition en k (k fixé) classes d'un ensemble de n individus. C'est un algorithme itératif.
Soit *I* une population d'individus, cette population est représentable sur R et forme un nuage de n points.

On cherche à constituer une partition en k classes sur i. chaque classe est représentée par son centre, également appelé noyau, constitué du petit sous-ensemble de la classe qui minimise le critère de dissemblance. [12]

B. La classification hiérarchique

La classification hiérarchique : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus

élevé, ils seront distingués et appartiendront à deux sous- groupes différents. Le résultat d'une classification hiérarchique n'est pas une partition de l'ensemble des individus. C'est une hiérarchie de classes telle que :

-Toute classe est non vide.

-Tout individu appartient à une (et même plusieurs) classes.

-Deux classes distinctes sont disjointes, ou vérifient une relation d'inclusion (l'une d'elle est incluse dans l'autre)

-Toute classe est la réunion des classes qui sont incluses dans elle.

L'avantage de cette méthode est qu'elle n'est soumise à aucune initialisation particulière de paramètre(s) ce qui la rend déterministe, et en outre, que le nombre de classes n'a pas à être fixé a priori. Cependant, ce type de méthode impose le calcul de la matrice des distances de tous les points d'observation avec tous les autres, et cette masse de calculs est beaucoup trop importante compte tenu du temps que nous voulons consacrer à cette étape. [11]

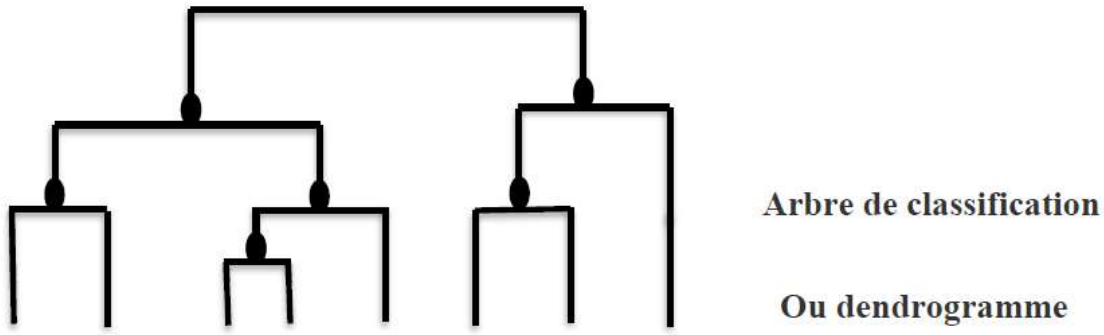

Figure2: Classification selon la méthode hiérarchique.

Parmi les méthodes non-supervisées les plus utilisées, citons deux types d'approches :

La classification hiérarchique ascendante

La CAH permet de construire une hiérarchie entière des objets sous la forme d'un "arbre" dans un ordre ascendant. On commence en considérant chaque individu comme

une classe et on essaye de fusionner deux ou plusieurs classes appropriées (selon une similarité) pour former une nouvelle classe. Le processus est étiré jusqu'à ce que tous les individus se trouvent dans une même classe. Cette classification génère un arbre que l'on peut couper à différents niveaux pour obtenir un nombre de classes plus ou moins grand.

Différentes mesures de la distance inter classes peuvent être utilisées : la distance euclidienne, la distance inférieure (qui favorise la création de classes de faible inertie) ou la distance supérieure (qui favorise la création de classes d'inertie plus importante) etc.

Dans le cas de la classification ascendante hiérarchique, à partir des éléments, on forme des petites classes ne comprenant que des individus très semblables, puis à partir de celle-ci, on construit des classes de moins en moins homogènes, jusqu'à obtenir la classe tout entière. [13]

Le principe de CAH :

La classification ascendante hiérarchique (CAH) est une méthode de classification itérative dont le principe est simple :

Prenons un échantillon de données et apprenons comment fonctionne étape par étape.

- Tout d'abord, faites de chaque point de données un « cluster unique », qui forme N clusters. (supposons qu'il y a N nombre de clusters).

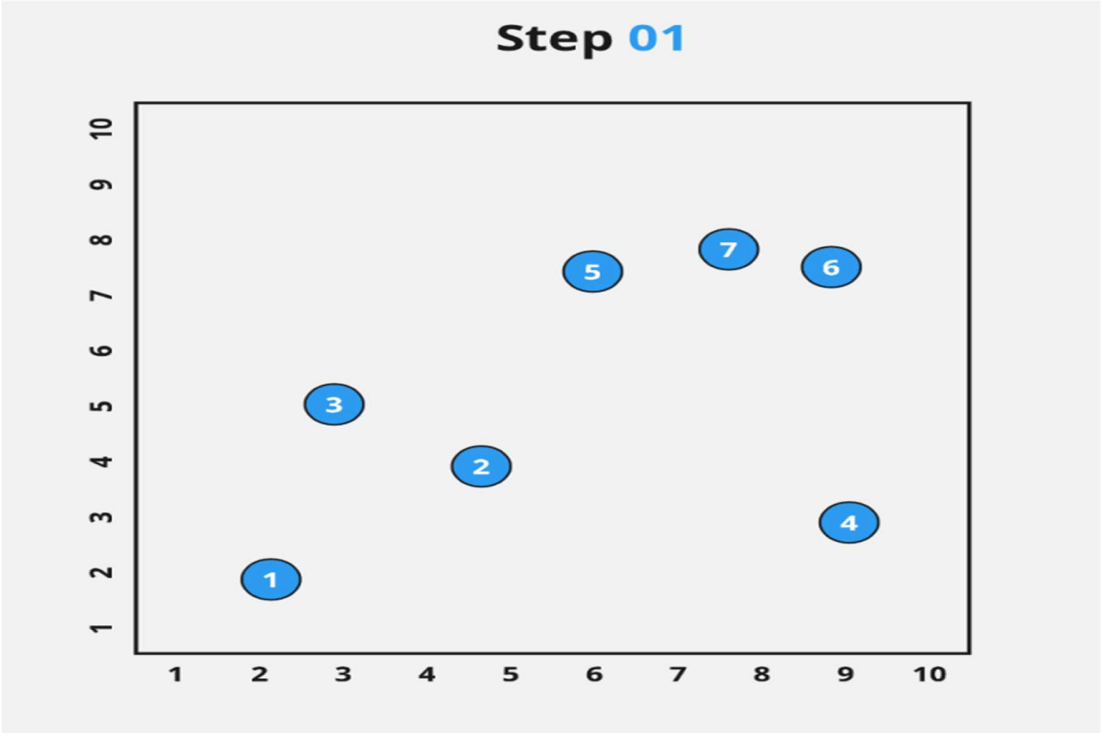

Figure 3 : La première étape de CAH

- Prenez les deux prochains points de données les plus proches et faites-en un seul cluster ; maintenant, il forme des grappes N-1.

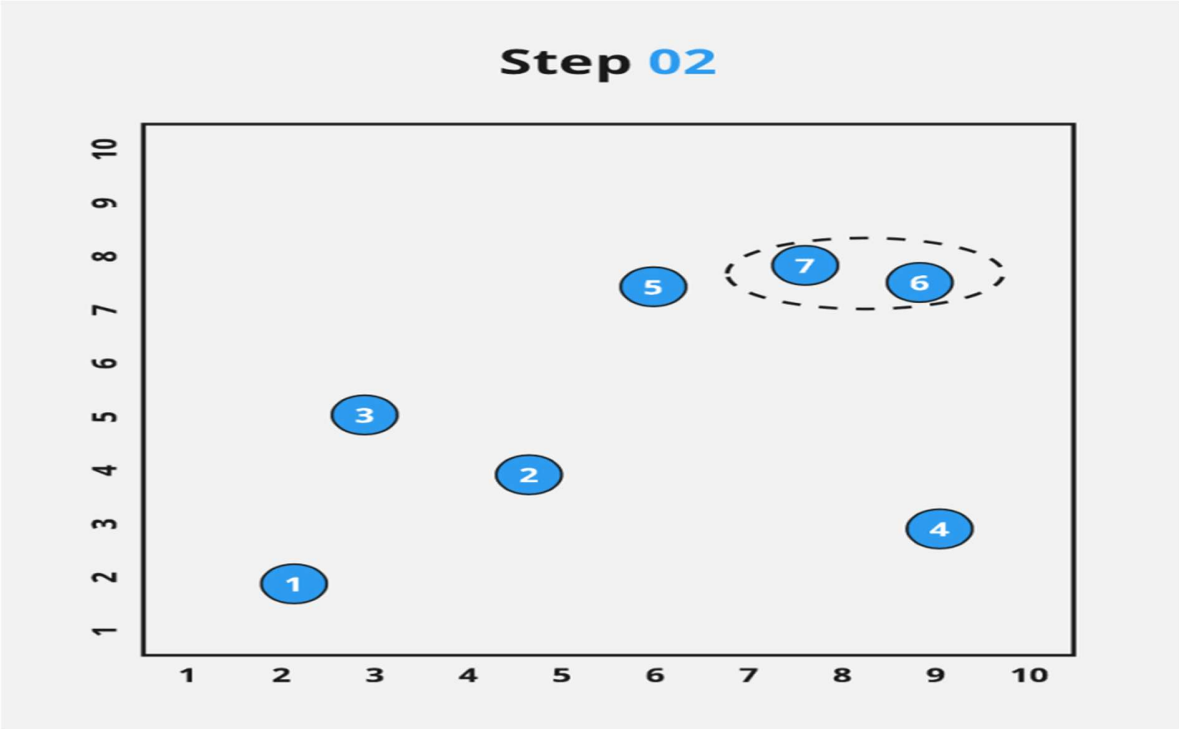

Figure 4 : La deuxième étape de CAH

- Encore une fois, prenez les deux clusters et faites-en un seul cluster ; maintenant, il forme N-2 clusters

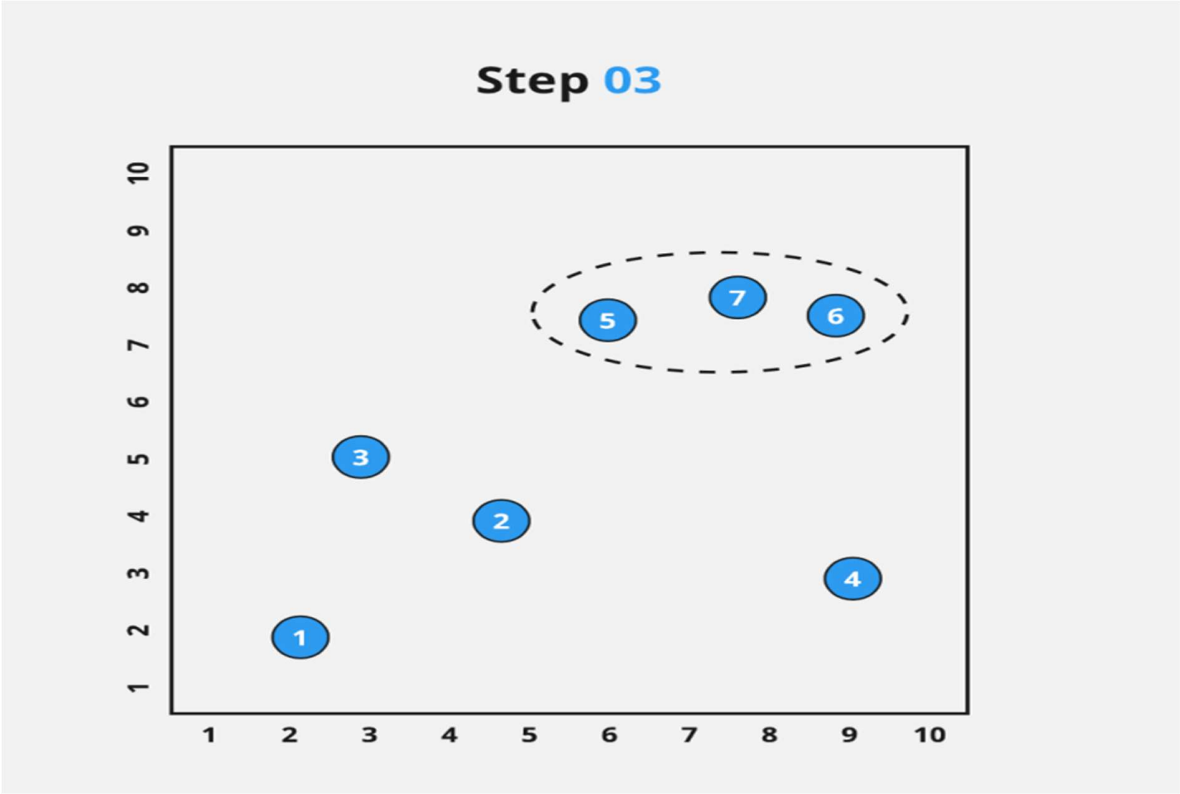

Figure 5 : La troisième étape de CAH

- Répétez « Step 3 » jusqu'à ce qu'il ne vous reste qu'un seul cluster

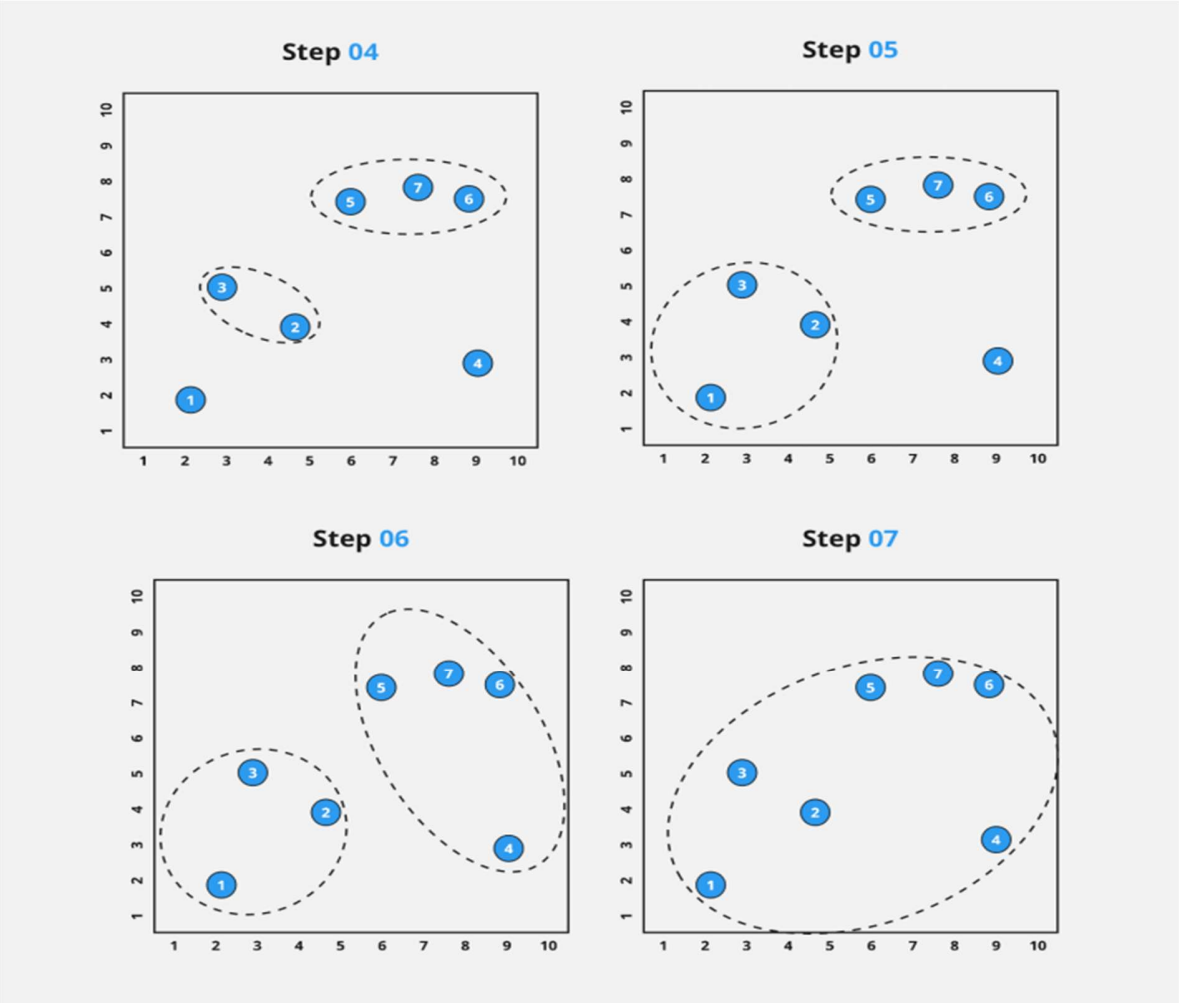

Figure 6: Différents étapes jusqu'aux résultats final [15]

Ces regroupements successifs produisent un arbre binaire de classification (dendrogramme), dont la racine correspond à la classe regroupant l'ensemble des individus. Ce dendrogramme représente une hiérarchie de partitions. On peut alors choisir une partition en tronquant l'arbre à un niveau donné, le niveau dépendant soit des contraintes de l'utilisateur (l'utilisateur sait combien de classes il veut obtenir), soit de critères plus objectifs. [15]

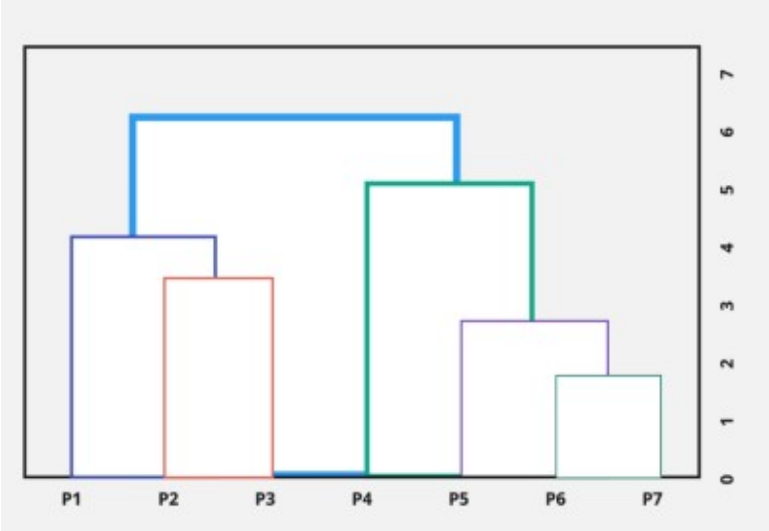

Figure 7: Dendrogramme représente résultat final. [15]

Les avantages :

La (CAH) est une méthode de classification qui présente les avantages suivants :

- On travaille à partir des dissimilarités entre les objets que l'on veut regrouper. On peut donc choisir un type de dissimilarité adapté au sujet étudié et à la nature des données.
- L'un des résultats est le dendrogramme, qui permet de visualiser le regroupement progressif des données. On peut alors se faire une idée d'un nombre adéquat de classes dans lesquelles les données peuvent être regroupées.
- L'algorithme du « CHA » en général consiste à fournir un ensemble de partitions de moins en moins fines obtenues par regroupement successifs de parties.

La classification hiérarchique descendante

Dans la CDH, on considère tous les individus comme une seule classe au début, on divise successivement les classes en classes plus raffinées. Le processus marche jusqu'à ce que chaque classe contienne un seul point ou bien si l'on atteint un nombre de classes désiré. [13]

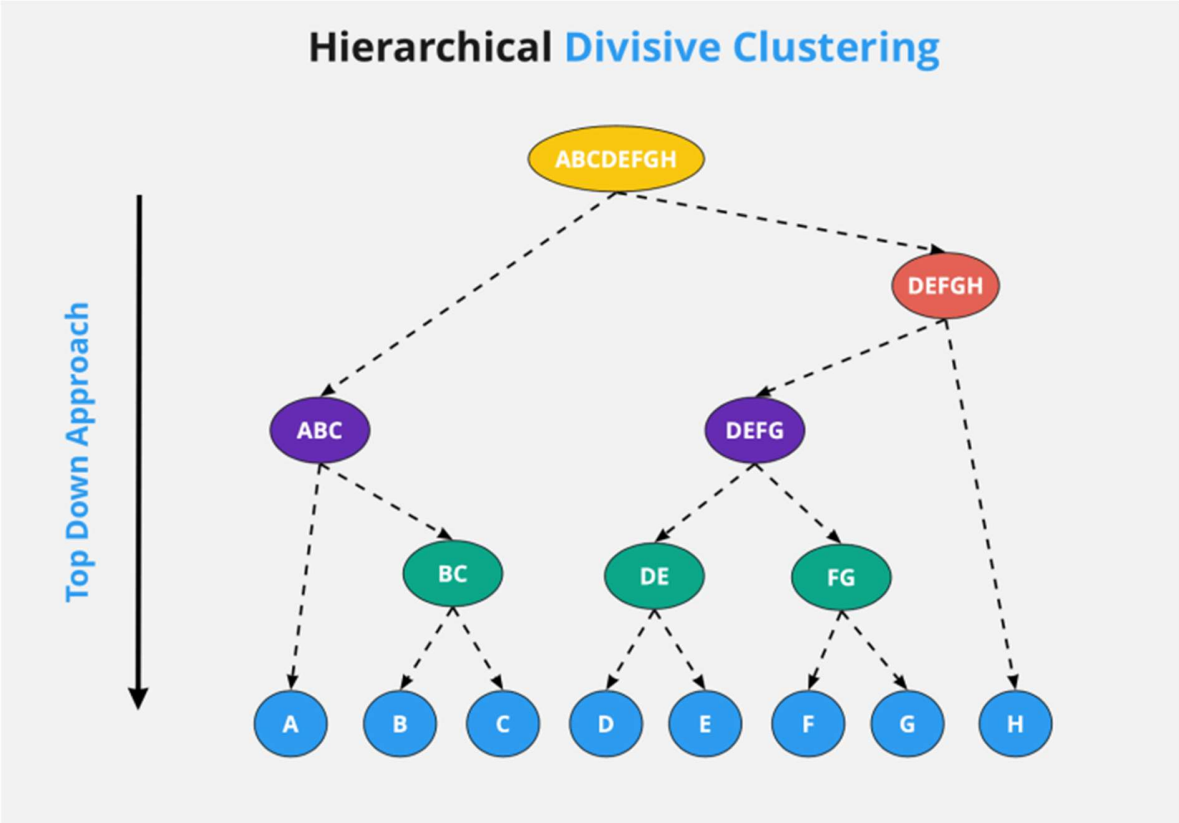

Figure 8: Schémas de la classification descendante.

4.Conclusion

Dans ce chapitre, nous introduisons la généralité des bases de données et les différents types existants, puis continuons à introduire la classification des bases de données et les différentes méthodes de classification.

Dans le chapitre suivant, nous présenterons la classification hiérarchique ascendante et ses avantages, puis comparerons et étudierons différentes méthodes d'agrégation.

CHAPITRE II : APPROCHES D'AGREGATIONS

1 Introduction

Pour classer des données de quelle que soit leur nature, nous appliquant le clustering Hiérarchique, ce qui est un regroupement d'objets similaires selon un tel critère, ce regroupement peut être effectué d'une manière ascendante. Cette manière avait des méthodes de calcul la distance entre les classes.

2 Les différentes approches ou stratégie d'agrégation :

Il existe plusieurs façons de mesurer la distance entre les deux afin de décider des règles de regroupement, et elles sont souvent appelées méthodes de linkage.

Certaines des méthodes de liaison populaires sont:

- Single Linkage
- Complete Linkage
- Average Linkage
- Centroid Linkage
- Ward's Linkage

A. L'algorithme Single-Linkage:

La distance entre deux clusters est représentée par la distance minimale entre toutes les paires de données entre les deux clusters (paire composé d'un élément de chaque cluster), nous parlons alors de saut minimum.

Le point fort de cette approche est qu'elle sait très bien détecter les classes allongées, mais son point faible est qu'elle est sensible à l'effet de chaîne, [Tufféry, 2005] et donc moins adaptées pour détecter les classes sphériques. [16]

Chaine : Nous appelons effet de chaîne lorsque deux points très éloignés l'un de l'autre mais reliés par une suite de points très proches les uns des autres sont rassemblés dans la même classe.

La formule de la distance maximale est :

$$\Delta(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{min}_{\boldsymbol{i}\in\boldsymbol{A},\boldsymbol{j}\in\boldsymbol{B}}\boldsymbol{d}(\boldsymbol{i},\boldsymbol{j}) \ldots\ldots. (\mathbf{1})$$

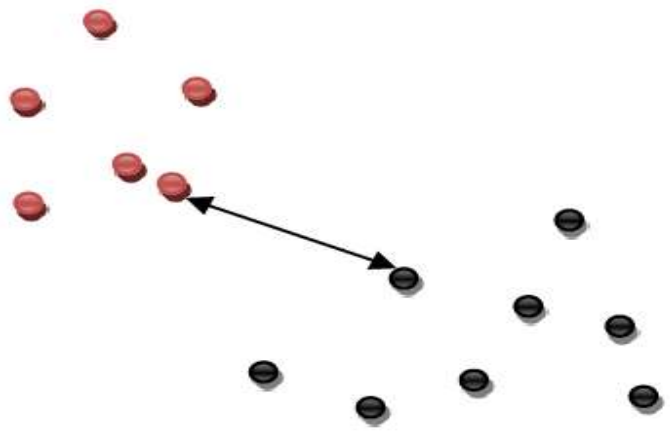

Figure 9: Schéma de la classification « Single-Linkage ».

B. L'algorithme Complete-linkage

La distance entre de clusters est représentée par la distance maximum entre toutes les paires de données entre les deux clusters, nous parlons alors de saut maximum ou de critère du diamètre.
Par définition cette approche est très sensible aux points aberrants donc elle est peu utilisée [Tufféry, 2005]. Et bien on démontre aussi le résultat de cet algorithme et comment il fonctionne. [17]

La formule de la distance maximale est :

$$\Delta(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{max}_{\boldsymbol{i}\in\boldsymbol{A},\boldsymbol{j}\in\boldsymbol{B}}\boldsymbol{d}(\boldsymbol{i},\boldsymbol{j}) \ldots\ldots (\mathbf{2})$$

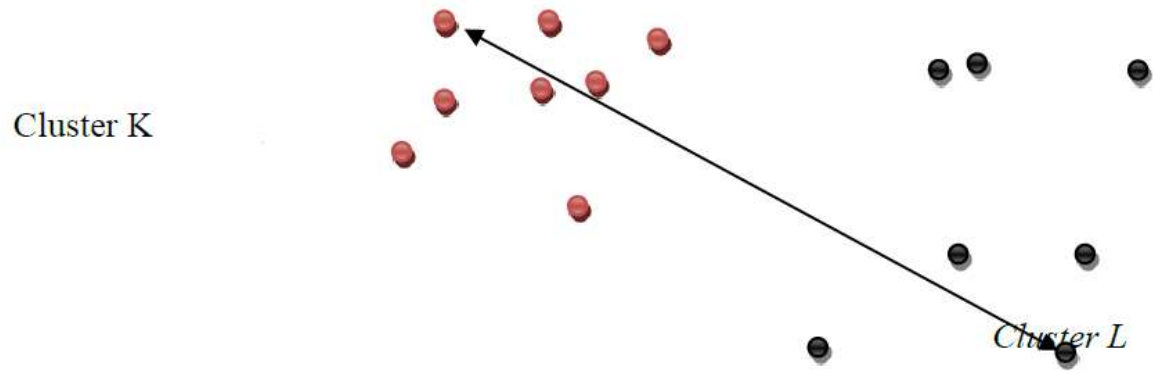

Figure 2: Schéma de la classification « Complete-Linkage ».

c. L'algorithme Average-Linkage

Propose de calculer la distance entre deux clusters en prenant la valeur moyenne des distances entre tous les couples d'objets des deux clusters. Nous parlons aussi de saut moyen. Cette approche tend à produire des classes de même variété.

$$\Delta(\boldsymbol{A},\boldsymbol{B}) = \frac{\mathbf{1}}{|\boldsymbol{A}|}\frac{\mathbf{1}}{|\boldsymbol{B}|}\sum\sum \boldsymbol{d}(\boldsymbol{i},\boldsymbol{j}) \dots\dots\dots (\mathbf{3})$$

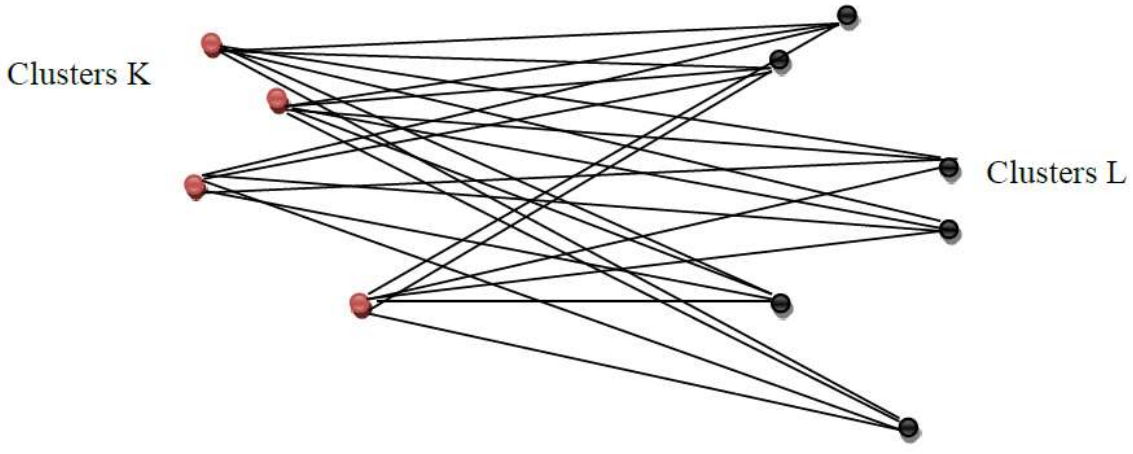

Figure 3: Schéma de la classification « Average-Linkage ».

D. L'algorithme Centroid-linkage (ou saut barycentrique)

C'est la distance entre deux clusters comme la distance entre leur centre de gravité. Une telle méthode est plus robuste aux points aberrants. Toutefois, elle est limitée aux données quantitatives numériques pour lesquelles le calcul du centre de gravité est possible.

Figure 42: Schéma de la classification «Centroid -Linkage ».

E. L'algorithme Ward-Linkage:

C'est la méthode la plus courante. Elle consiste à réunir les deux clusters dont le regroupement fera là moins baisser l'inertie interclasse. C'est la distance de Ward qui est utilisée pour : la distance entre deux classes est celle de leurs barycentres au carré, pondéré par les effectifs des deux clusters. On suppose tout de même l'existence de distances euclidiennes. [15]

- combinaison de clusters où l'augmentation de la variance intra-cluster est la plus faible.

-l'objectif est de minimiser la variance totale au sein de la grappe [15].

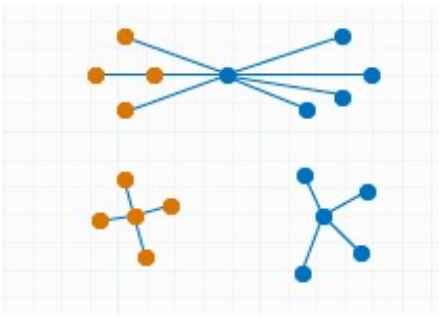

Figure 5: Schéma de la classification «Ward-Linkage ». [15]

Cette figure représente un schéma explicatif de méthode de Ward

Après le regroupage de chaque clusters la méthode de Ward commence le part 2 par sélectionné le centre de chaque cluster puis elle calcule le centre pour les 2 clusters dans une la même point.

3 Comparaison générale et discussion

Le tableau suivant récapitule les approches vues avec une comparaison entre ces dernières :

Tableau 1: Comparaison entre les différents algorithmes de CAH.

Algorithme	Avantage	Inconvénient
Single-Linkage	-Les algorithmes de liaison unique sont les meilleurs pour capturer des clusters de différentes tailles.	-ne peuvent pas regrouper correctement les clusters s'il y a du bruit (valeurs aberrantes) entre les clusters
Complete-linkage	-réussit également à séparer les clusters s'il y a du bruit entre les clusters.	-ont tendance à casser les grands clusters. -est biaisée vers les clusters globaux.
Average-Linkage	-réussit également bien à séparer les clusters s'il y a du bruit entre les clusters.	-est biaisée vers les clusters globaux.
Centroid-linkage **//avg+completlink**	Linkage réussit également à séparer les clusters s'il y a du bruit entre les clusters.	-est biaisée vers les clusters globaux.
Ward-Linkage	-produit généralement de meilleures hiérarchies de clusters	-est biaisée vers les clusters globaux.

	-est moins sensible au bruit (valeurs aberrantes).	

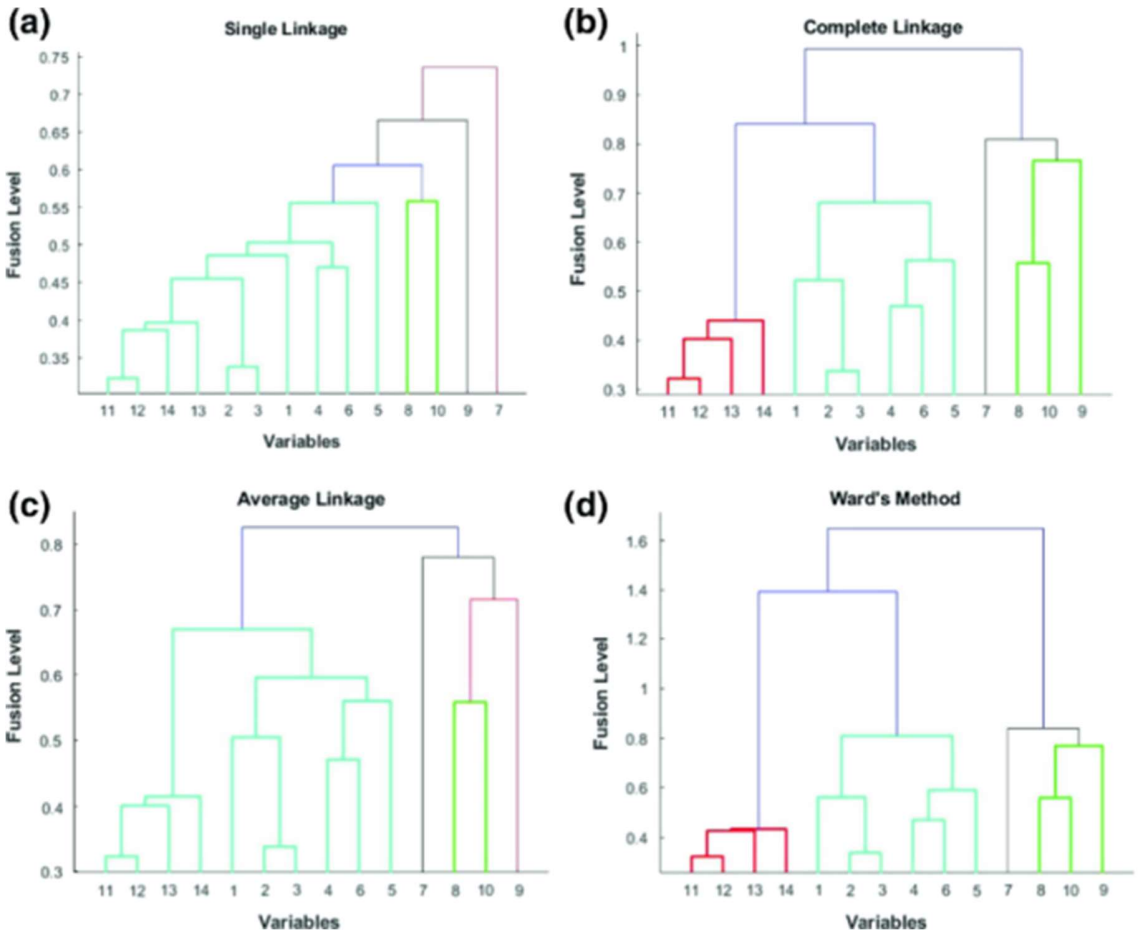

Figure 6: La différence entre les quatre méthodes dans le résultat final (dendrogramme de chaque algorithme). [15]

3 Conclusion :

D'après l'étude et l'analyse des différentes méthodes de classification, on a pu constater que la classification ascendante hiérarchique est celle qui offre plus d'avantages que les autres méthodes, elle fonctionne en recherchant à chaque étape les classes les plus proches pour les fusionner, l'étape la plus importante dans l'algorithme réside dans le choix de la distance entre deux classes et l'approche la plus connaît de calcul la distance est la méthode de Ward pour cela on choisit la meilleure méthode pour hiérarchiser de clusters.

PARTIE II : CONCEPTION

1 Introduction

Dans la partie précédente nous avons étudié le clustering et l'approches CAH (Classification Ascendante hiérarchique), et nous avons comparé les différentes solutions proposées dans le calcul de distance entre les clusters. Cette étude nous facilite la réalisation de notre solution qui va être présentée dans ce chapitre.

Nous allons donc dans cette partie à présenter la phase conceptuelle de la solution, nous commençons par présenter l'algorithme de Ward puis l'organigramme et les différentes fonctions qui ont utilisé dans la partie conception.

2 L'algorithme Ward-Linkage

2.1 Introduction

Si on peut considérer E comme un nuage d'un espace, on agrège les individus qui font le moins varier l'inertie intra-classe. A chaque pas, on cherche à obtenir un minimum local de l'inertie intra-classe ou un maximum de l'inertie interclasse.
L'indice de dissimilarité entre deux classes (ou niveau d'agrégation de ces deux classes) est alors égal à la perte d'inertie-classe résultant de leur regroupement ; Le regroupement des données soit des points ou autre démarre d'une façon hiérarchique aléatoire et finisse en donnant l'arbre avec ses différents niveaux. [18]

2.2 Définition

La méthode de Ward (1987) offre une approche itérative à la construction de consensus qui encourage le développement et la considération du point de vue unique de chaque contributeur. Les collaborateurs commencent par s'orienter vers la méthode et des buts spécifiques au projet, puis s'engagent dans un processus itératif, cyclique entre le travail créatif individuel et des entrevues de groupe. Les entrevues servent d'opportunités pour partager des idées dans une atmosphère non-critique. Paradoxalement, les plus grands partis du travail d'atteinte de consensus se passent pendant que les collaborateurs travaillent indépendamment ; au travers des itérations, les versions tendent à converger alors que les collaborateurs adaptent et adoptent ce qu'ils aiment des idées des autres.

Cette technique tend à regrouper les ensembles représentant les petites classes.
On calcule cette inertie entre les classes :

G_A : Centre de gravité de la classe A (PoidsP_A)

G_B : Centre de gravité de la classe B (PoidsP_B)

G_{AB} : Centre de gravité de leur réunion.

$$\boldsymbol{G_{AB} = \frac{P_A G_A + P_B G_B}{P_A + P_B}}$$

L'inertie interclasse étant la moyenne des carrés des distances des centres de gravité des classes au centre de gravité total, la variation d'inertie interclasse, lors du regroupement de A et B est égale à :

$$\mathbf{P_A}\boldsymbol{d^2}(\mathbf{G_A}, \boldsymbol{G}) + \mathbf{P_B}\boldsymbol{d^2}(\mathbf{G_A}, \boldsymbol{G}) - (\mathbf{P_A} + \mathbf{P_B})\boldsymbol{d^2}(\mathbf{G_{AB}}, \boldsymbol{g}) \ldots\ldots\ldots (\mathbf{4})$$

Elle vaut :

$$\Delta(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{P_A G_A + P_B G_B}}{\boldsymbol{P_A + P_B}} \boldsymbol{d^2}(\mathbf{G_A}, \mathbf{G_B})$$

A chaque itération, on agrège de manière à avoir un gain minimum d'inertie intra-classe

Δ (A, B) = perte d'inertie intra-classe due à cette agrégation. [18]

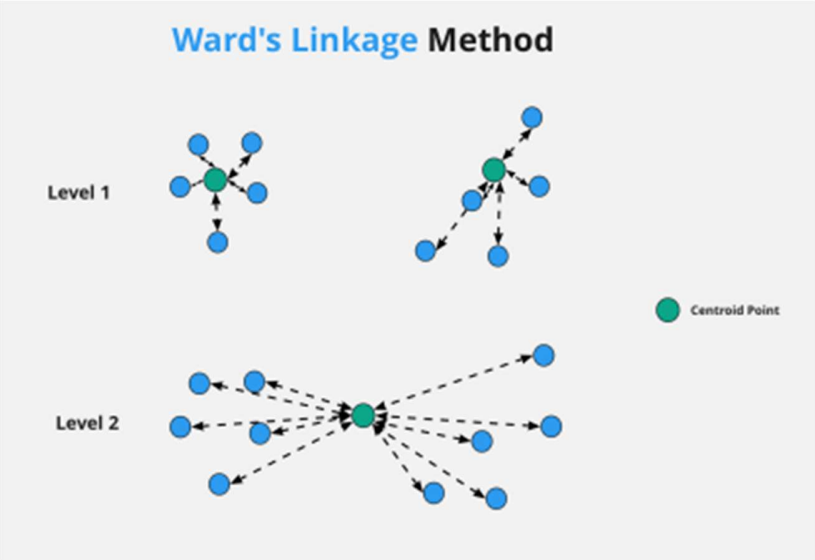

Figure 75: Méthode de Ward .[15]

Cette figure représente un schéma explicatif de méthode de Ward

Après le regroupage de chaque clusters la méthode de Ward commence le part 1 par sélectionné le centre de chaque cluster puis elle calcule le centre pour les 2 clusters dans une la même point.

2.3 Principe

L'algorithme de Ward

Entrée : Une base d'exemples S Une hiérarchie H contenant |S| clusters **Sortie :** Une hiérarchie H mise à jours

Etapes : l'utilisateur va suivre ces étapes jusqu'à obtenir la sortie désirée

Etablir la table TD (table des distances) des valeurs de D(x,y) x et $y \in S$ parcourant S.

Tant que la table TD à plus d'une colonne (>1) faire

-Choisir les deux sous-ensembles H_i , H_j de S tels que D (H_i , H_j) est la plus petite.

-Supprimer H_j de la table, et remplacer H_i par $H_i \cup H_j$ (pour les colonnes et lignes).

-Ajouter un nouveau cluster dans la hiérarchie H dont les fils sont H_i etH_j.

-Calculer les distances de Ward entre $H_i \cup H_j$ et les autres éléments de la table. Et mettre à jour la table.

Fin tant que.

Retourner H

2.4 Organigramme d'algorithme de clustering à base de Ward

D'un point de vu générale l'algorithme consiste à classifier les points via une façon Hiérarchique.

L'entrée de l'algorithme est une base d'exemples avec une hiérarchie initiale des clusters des exemples. Quand les exemples (points) sont donnés l'algorithme va calculer la table des distances.

L'étape suivante ou il va calculer l'inertie minimale entre les centres de gravitée de chaque point représenté comme vecteur dans la TD initial et les grouper selon ce critère puis inclure ce groupe de deux dans la hiérarchie.

Et effet l'algorithme utilise à chaque fois les points (les exemples en entrée) d'après la Table des Distances alors cette dernière va être mise à jours ainsi que l a hiérarchie.
L'algorithme va s'arrêter la faite qu'il va rester dans la Table des Distances (TD) qu'un vecteur qui représente un point puisqu'il faut avoir au plus un vecteur. [19]
La hiérarchie finale est obtenue sous forme d'arbre.

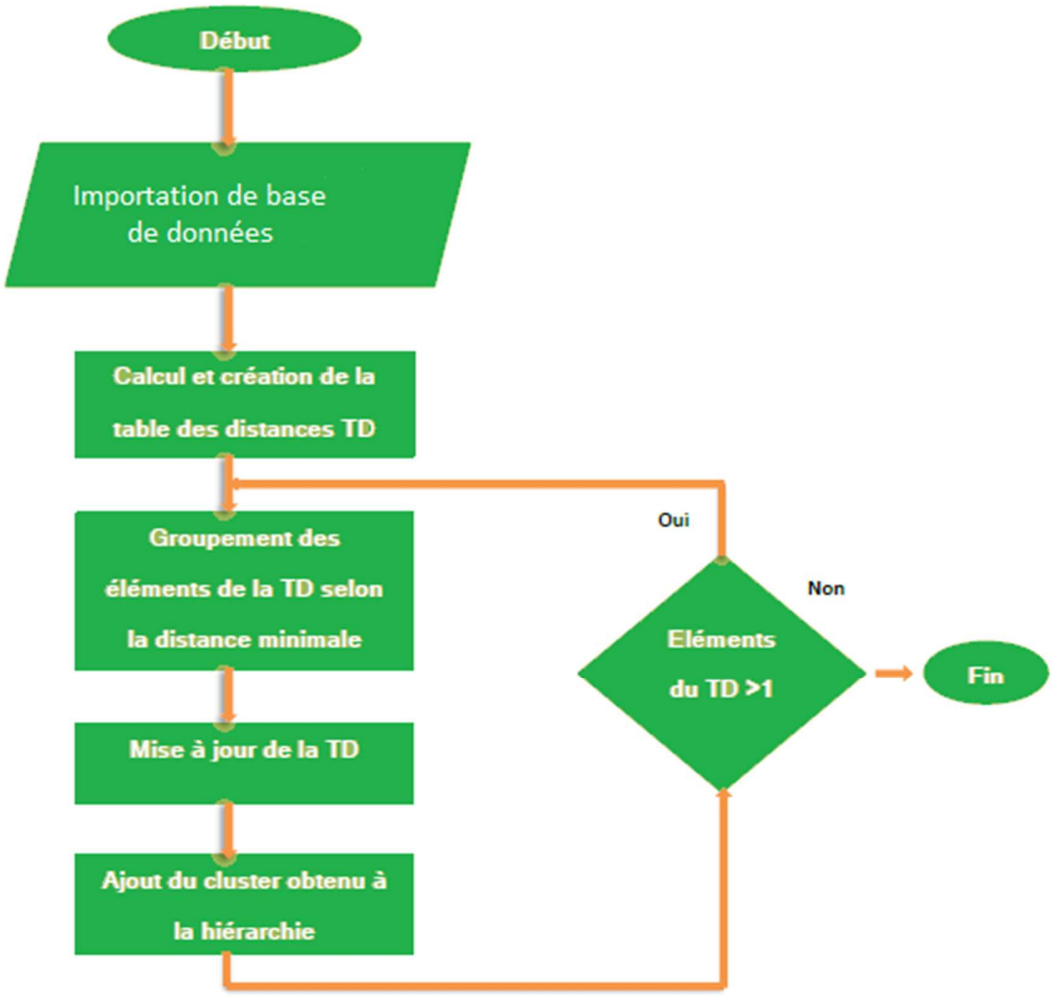

Figure 8: Organigramme de l'algorithme de Ward.

L'objectif de l'algorithme est de fournir une hiérarchie de nœuds, chacun d'eux contient les exemples appartenant au même cluster.

2.5 Fonction

AgglomerativeClustering

Fonctionne de manière « bottom-up ». C'est-à-dire que chaque objet est initialement considéré comme un cluster à un seul élément (feuille). A chaque étape de l'algorithme, les deux clusters les plus similaires sont combinés en un nouveau cluster plus grand (nœuds). Cette procédure est itérée jusqu'à ce que tous les points soient membres d'un seul grand cluster (racine). [19]

3 Conclusion

Dans cette partie, nous avons présenté les différentes parties et étapes de notre système. Surtout, nous approfondissons dans chaque partie pour mieux comprendre les bases de toutes les méthodes utilisées afin que tout soit clair.

Dans la partie suivante, nous présenterons nos expérimentations réalisées dans l'objectif d'évaluer de Ward méthode. Nous décrirons tous les outils matériels (Hardware) et logiciels (Software) que nous avons utilisés pour construire notre système. De plus, nous présenterons les résultats que nous avons obtenus dans notre système.

PART III : IMPLEMENTATION ET EVALUATION

1 Introduction

Après l'étape de modélisation de notre approche, nous arrivons dans cette dernière partie à l'implémentation et la mise en œuvre de la solution proposée, et nous présenterons les résultats que nous avons obtenus dans notre système.

Pour réaliser nos recherches et construire notre système de classification ascendante hiérarchique, nous avons utilisé un ensemble d'outils de logiciels et de matériels.

2 Matériel informatique (Hardware)

Pour réaliser notre système, nous avons utilisé un ordinateur qui tourne avec un système d'exploitation Windows 7, alimenté par un processeur Intel® Core™ i5-6200U 2.3GHz x64 avec une carte graphique Intel® HD Graphics 520, et une RAM de 8 GO.

3 Logiciel (Software)

Pour construire notre modèle, nous utilisons Jupyter, un environnement de travail simple et facile à utiliser.

De plus, nous avons utilisé Python 3.7 comme langage de programmation. Nous avons déduit des travaux antérieurs qu'il s'agit du langage le plus approprié pour de nombreuses raisons :

- C'est un langage interprété, ce qui signifie que Python exécute directement le code ligne par ligne.
- Un langage de programmation de haut niveau a une syntaxe de type anglais. Il est plus facile de lire et de comprendre le code.
- C'est un langage très large, il est utilisé pour tous les domaines.
- Il dispose de bibliothèques diverses et riches, celles dédiées au deep learning et celles utilisées pour la gestion d'autres structures de données et autres.
- Sa simplicité permet d'exprimer des équations et des formules complexes, c'est pourquoi il est facile, ce qui consomme moins de temps pour l'ensemble du processus de développement.

4 Les Bibliothèque

Dans cette partie, nous présenterons toutes les bibliothèques utilisées pour construire notre modèle.

Matplotlib est une bibliothèque du langage de programmation Python destinée à tracer et visualiser des données sous formes de graphiques. Elle peut être combinée avec les bibliothèques python de calcul scientifique NumPy et SciPy. Matplotlib est distribuée librement et gratuitement sous une licence de style BSD. Sa version stable actuelle (la 2.0.1 en 2017) est compatible avec la version 3 de Python. [20]

SciPy est un projet visant à unifier et fédérer un ensemble de bibliothèques Python à usage scientifique. Scipy utilise les tableaux et matrices du module NumPy . [21]

Pandas : est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données. Elle propose en particulier des structures de données et des opérations de manipulation de tableaux numériques et de séries temporelles. [22]

Scikit-learn : est une bibliothèque Python gratuite et open source, créée par plusieurs chercheurs pour l'apprentissage automatique et l'apprentissage en profondeur, et elle est commercialement utilisable. Il permet le passage à des versions efficaces de nombreux algorithmes actuels. [23]

NumPy : c'est l'une des nombreuses bibliothèques open source de python, c'est pour les tableaux numériques. Il a été créé spécialement pour le calcul scientifique, en particulier le calcul matriciel, tout en offrant de multiples fonctions permettant la création et la manipulation de matrices et de vecteurs. [21]

5 Base de données (Dataset)

Dans cette section, nous discuterons de la base de données utilisée dans notre système.

CASBAH Foods cet ensemble de données se compose de critiques alimentaires de SARL CASBAH.

L'ensemble de données contient un fichier nommé "casbahdzqlsql.sql".

Les données contenues dans le fichier "casbahdzqlsql.sql" représentent 57 produits et 3417 commandes.

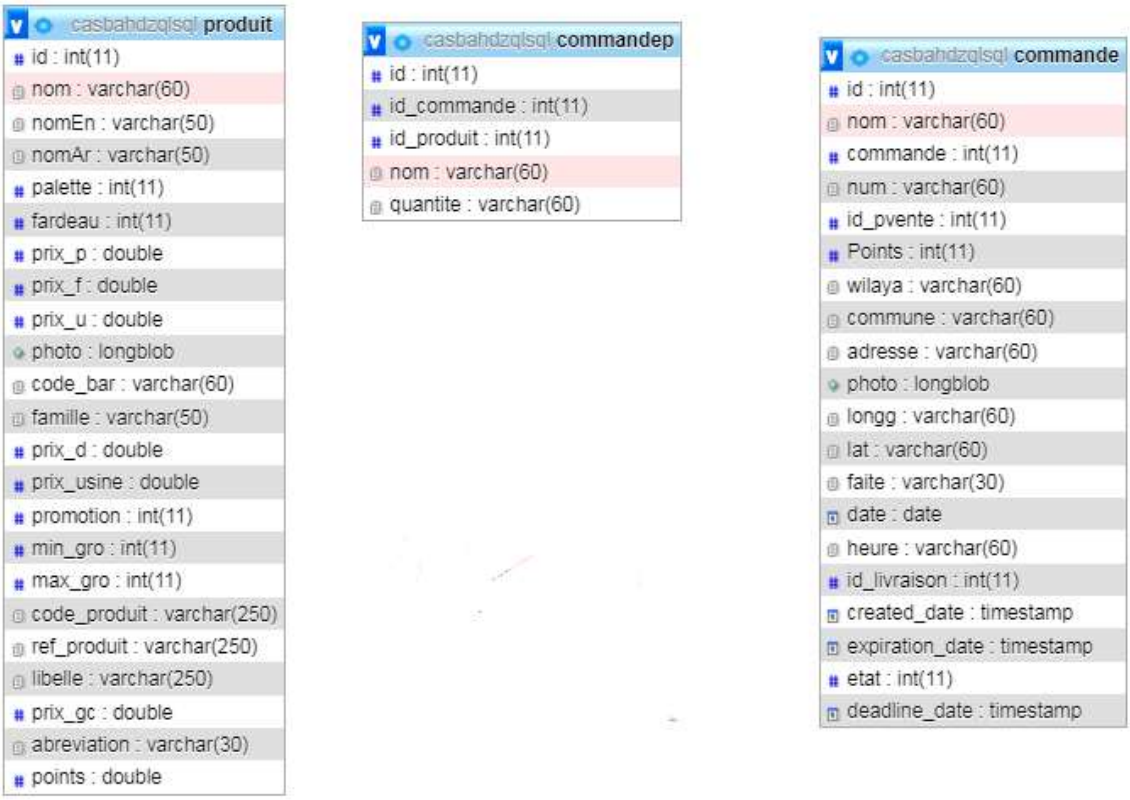

Figure 9: Représente les importantes tables dans notre système.

Après le choisir des tables importantes dans notre dataset on crée une vue.

```
1 CREATE VIEW region AS
2 SELECT * FROM commandep,commande,produit
3 WHERE commandep.id_produit=produit.id
4 AND commandep.id_commande=commande.id
```

Figure 10: Représente la commande SQL pour créer la vue.

6 Implémentation :

Dans cette section, nous allons commencer le travail principal pour construire notre système à l'aide de la méthode de Ward. Les données sont divisées en des catégories selon des différents critères (Régions…) pour faciliter l'utilisation dans la classification des bases de données.

Chaque catégorie est représentée dans une vue dans notre base des données, on exporte chaque vue à un fichier avec une extension .csv

Après la division en catégories, nous avons utilisé un fichier s'appelle « commande.csv » qui est une partie du notre datas et comme input dans notre système. Puis nous avons utilisé l'approche « Ward » pour calculer la distance entre les points de données les plus proches l'un de l'autre, cette approche nous permet de minimiser la variance entre les clusters on regroupe les points de données les plus proches dans un même cluster ainsi de suite jusqu'où nous obtenons un seul cluster. Encore nous avons utilisé la méthode ''fit_predict'' qui permet de renvoyer les noms des clusters auxquels appartient chaque point de données.

A la fin nous allons présenter notre résultat par un dendrogramme.

7 La classification de base de données

Nous avons utilisé la méthode Ward pour calculer la distance entre deux points de données les plus proches entre eux et faire le clustering dans le but de classifier notre base de données.

La formule mathématique suivante nous permet de calculer la distance entre les deux clusters :

$$\Delta(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{P_A G_A + P_B G_B}}{\boldsymbol{P_A + P_B}} \boldsymbol{d^2(G_A, G_B)}$$

La figure suivante représente la méthode Ward-linkage:

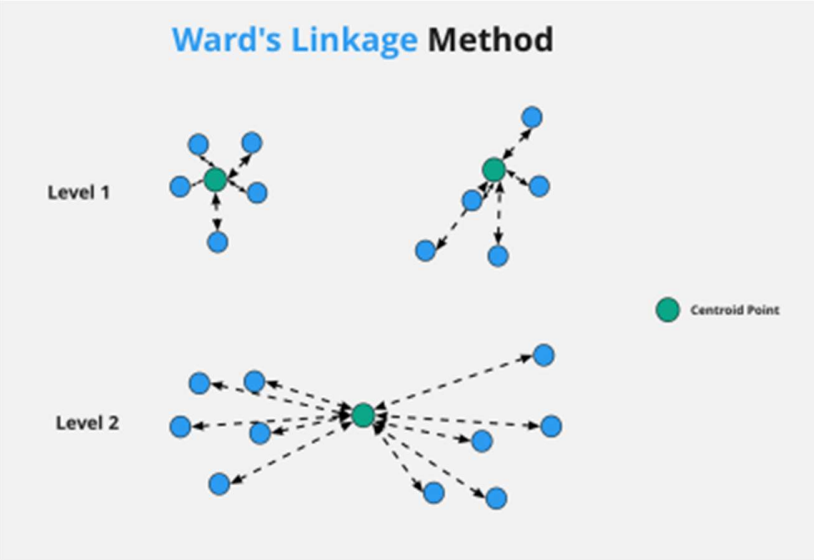

Figure 19: La méthode Ward-linkage [14].

8 Code source

Maintenant, nous expliquons une partie de notre code source que nous avons utilisée pour construire notre système.

```
Entrée [ ]: import matplotlib.pyplot as plt
            import pandas as pd
            %matplotlib inline
            import numpy as np
```

Figure 20: Importation des bibliothèques

Ensuite, pour importer l'ensemble de données, exécutez le code suivant :

```
Entrée [ ]: customer_data = pd.read_csv('commande.csv')
```

Figure 11: Importation de base des données

Vérifier le nombre d'attributs :

```
Entrée [11]: customer_data.shape
```

Figure 122: Vérification le nombre d'attributs

```
Entrée [15]: customer_data.columns=['id','id_pvente','etat']
```

Figure 13: Commande de choisir des colonnes

Nous utiliserons à nouveau la bibliothèque scipy pour créer les dendrogrammes de notre ensemble de données. Exécutez le script suivant pour ce faire :

```
Entrée [14]: import scipy.cluster.hierarchy as shc

             plt.figure(figsize=(10, 7))
             plt.title("Customer Dendograms")
             dend = shc.dendrogram(shc.linkage(data, method='ward'))
```

Figure 14:Commande pour applique la méthode de Ward et affiche le dendrogramme

La figure ci-dessus, nous importons la classe de hiérarchie de la bibliothèque scipy.cluster en tant que shc. La classe de hiérarchie a une méthode de dendrogramme qui prend la valeur retournée par la méthode de liaison de la même classe. La méthode de liaison prend l'ensemble de données et la méthode pour minimiser les distances comme paramètres. Nous utilisons 'Ward' comme méthode car elle minimise ensuite les variantes de distances entre les clusters.

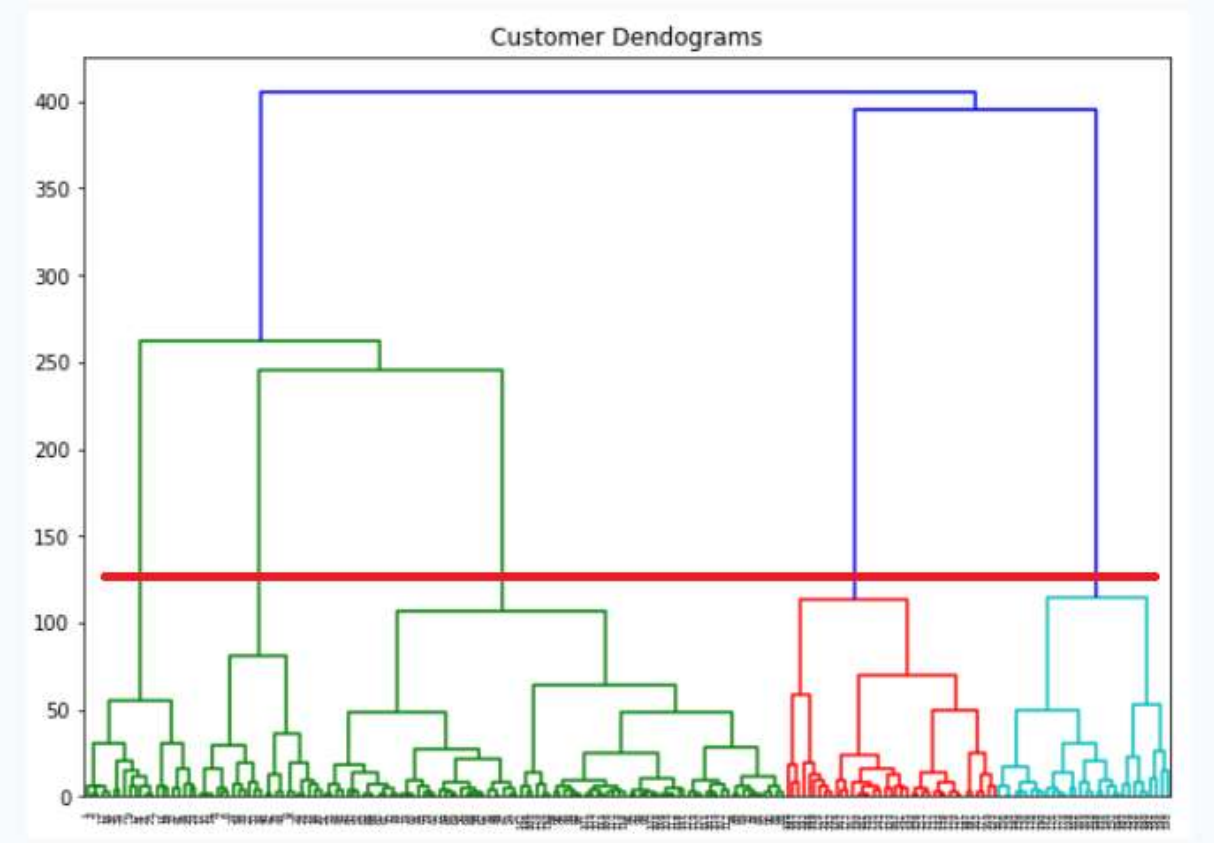

Figure 15: Dendrogramme représente le résultat de méthode de Ward

Maintenant que nous connaissons le nombre de clusters pour notre ensemble de données, l'étape suivante consiste à regrouper les points de données dans ces cinq clusters (dans notre cas en ont 5 clusters). Pour ce faire, nous utiliserons à nouveau la classe AgglomerativeClustering de la bibliothèque sklearn.cluster.

```
Entrée [15]: from sklearn.cluster import AgglomerativeClustering

cluster = AgglomerativeClustering(n_clusters=5, affinity='euclidean', linkage='ward')
cluster.fit_predict(data)
```

Figure 16 : Commande pour affiche le résultat des points de données

Vous pouvez voir les étiquettes de clusters de tous vos points de données. Comme nous avions cinq clusters, nous avons cinq étiquettes dans la sortie, c'est-à-dire 0 à 4.

Comme dernière étape, traçons les clusters pour voir comment nos données ont été regroupées :

```
Entrée [16]: plt.figure(figsize=(10, 7))
             plt.scatter(data[:,0], data[:,1], c=cluster.labels_, cmap='rainbow')
```

Figure 17: Tracer les clusters.

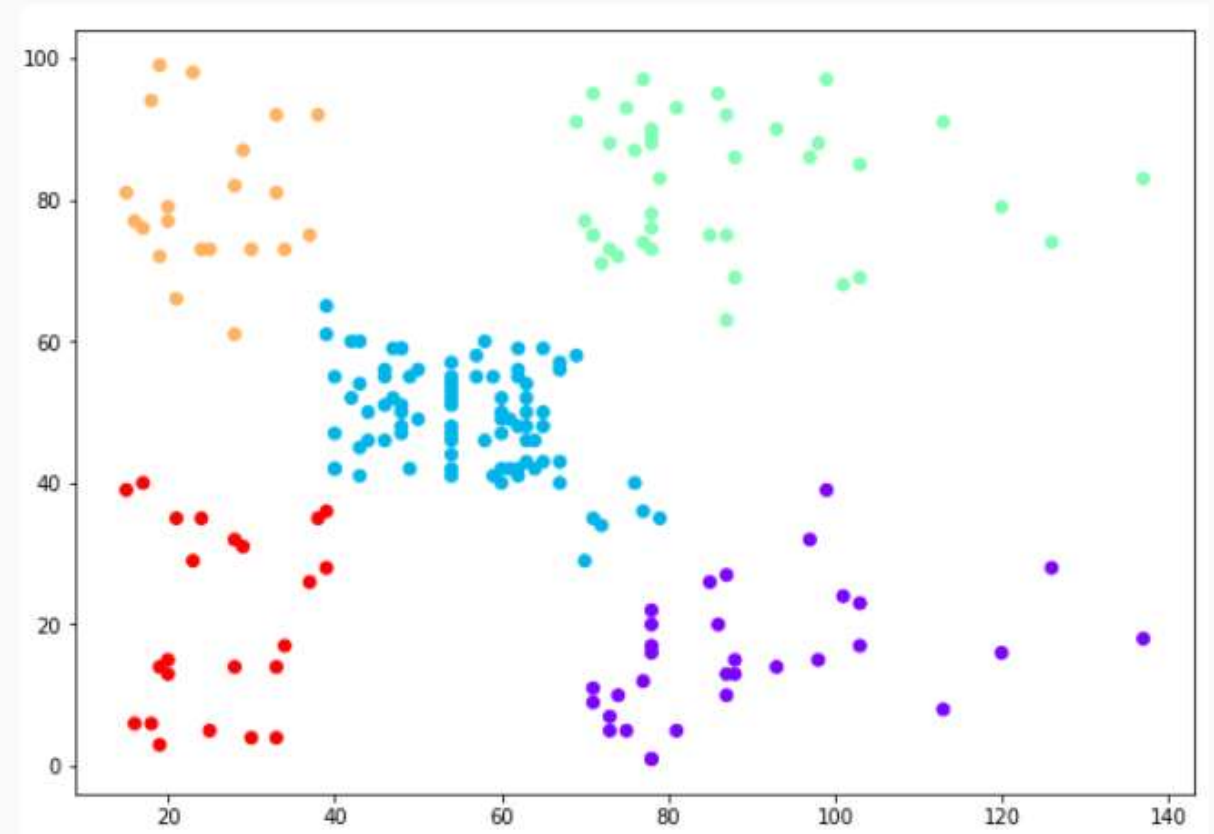

Figure 28: Résultat finale représente par des points de données.

9 Conclusion

Dans cette dernière partie, nous avons présenté et implémenté notre solution avec la classification ascendante hiérarchique base sur la méthode de Ward.

Nous avons aussi testé notre solution du notre problème et nous avons montré par les expérimentations de performance de notre solution qu'elle est capable de retourner de bons résultats de qui satisfont les besoins.

CONCLUSION GENERALE

La casbah est une société reçoit plusieurs commandes qui sont stockées dans la base de données mais cette dernière est devenue très volumineuse donc elle est très difficile à gérer et traiter par les administrateurs, donc l'entreprise souhaite apporter une solution afin de pouvoir gérer les commandes et les données massives.

Dans ce contexte, nous avons utilisé la classification ascendante hiérarchique à cause de ses différents avantages et points forts pour regrouper la base des données à différents clusters puis on utilise l'approche de Ward qui nous a permet de calculer la distance entre les deux clusters.

Nous avons utilisé Python comme langage de programmation et Jupyter comme environnement de travail pour développer et tester notre solution. Les expérimentations montrent que notre solution retourne de bons résultats de classification qui satisfont les besoins des administrateurs.

Cependant, notre solution laisse place à d'autres améliorations. Comme travaux futurs, nous pensons qu'il est possible d'optimiser notre système et d'utiliser plus de techniques pour obtenir de bons résultats de classification en optimisant le facteur temps.

Au niveau des difficultés, nous sommes confrontés à des problèmes de compréhension. Comme la plupart des documents qu'on pouvait nous prêter n'étaient pas nets et lucides, malgré les nombreuses heures passées devant l'ordinateur, les déplacements, les différents documents à consulter, etc…nous avons réellement eu beaucoup de plaisir à faire ce travail, et choisir ce thème.

BIBLIOGRAPHIE

[1] http://www.casbahdz.com/about_fr.html, dernière visite 23-10-2021.

[2] François-Xavier Jollois. Contribution de la classification automatique à la fouille de données. Ordinateur et société [cs.CY]. Université Paul Verlaine - Metz, 2003.

[3] Maurice Roux, Algorithmes de classification, Université Paul Cézanne, Marseille, France, (2006).

[4] Nakache, J.P, Confais, J. Approche pragmatique de la classification. Ed. Technip, Paris, (2005).

[5] Karem Fatma, Dhibi Mounir, and Martin Arnaud. Combinaison de classi_cation supervisée et non supervisée par la théorie des fonctions de croyance. International Conference on Belief Functions, Compiègne, France, mai 2012.

[6] E. Lebarbier, T. Mary-Huard Classification non supervisée, (2008).

[7] Lebart L., Morineau A., Piron M.,Statistique exploratoire multidimensionnelle, Dunod, 1997, 2004, 2006.

[8] Jodouin JF. Les réseaux de neurones- principe et définitions-. Paris: Hermes; (1994).

[9] Tan P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. (2014).

[10] E. E. Bron, M. Smits, W. J. Niessen and S. Klein, "Feature Selection Based on the SVM Weight Vector for Classification of Dementia," in IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 5, pp. 1617-1626, Sept. 2015, doi: 10.1109/JBHI.2015.2432832.

[11] (fr) Gilbert Saporta, Probabilités, Analyse des données et Statistiques, Paris, Editions Technip, (2006), 622.

[12] S. Mannor et al., “K-Means Clustering,” in Encyclopedia of Machine Learning, Boston, MA: Springer US, (2011), pp. 563–564.

[13] Samuel AMBAPOUR Introduction à l'analyse des données.

[14] V. Ridde and C. Dagenais, Approches et pratiques en évaluation de programmes. PU Montréal, (2012).

[15]https://dataaspirant.com/hierarchical-clustering-algorithm/#t-1608531820444, dernière visite 30-11-2021.

[16]Blum, A., Mitchell, T. combining labeled and unlabeled data with co-training. COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann, 1998, p. 92-100.

[17] Chevalier F. et Le Bellac J. La classification. Faculté des Sciences Économiques de Rennes, 2013, 44 p., disponible sur http://perso. univ-rennes1.fr/valerie.monbet/ExposesM2/2013/Classification2.pdf

[18] Pierre-Luis GONZALEZ, méthode de classification, (2008).

[19] Roux, M. (2018). A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. Journal of Classification.

[20] https://matplotlib.org/, dernière visite 30-11-2021.

[21] https://scipy.org/, dernière visite 30-11-2021.

[22] https://pandas.pydata.org/, dernière visite 30-11-2021.

[23] scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation. (2021). Scikit-Learn. https://scikit-learn.org/stable/, dernière visite 30-11-2021.