L'application de Data Mining à la Segmentation Client.
Une version améliorée de l'algorithme K-Means.

Mémoire présenté par

JOTI Anxhela

Pour l'obtention du Master 1 MIAGE

De l'université

Paris 1 Panthéon - Sorbonne

Année Universitaire :	2021 - 2022
Date de soutenance :	20 Juin 2022
Directeur de mémoire :	Daniel Diaz
Membre du jury :	Nicolas Herbaut

Remerciements

Je tiens à remercier l'équipe pédagogique de l'université Panthéon-Sorbonne pour leur professionnalisme et les connaissances académiques qu'ils m'ont apporté durant cette année de Master 1.

Je remercie également mon tuteur M. Daniel Diaz pour ses conseils et sa disponibilité. C'est grâce à son aide que j'ai pu développer un travail qualitatif.

Je souhaiterais exprimer ma reconnaissance envers mes camarades de la Miage Classique pour leur soutien moral et intellectuel tout au long de ma démarche.

Table des matières

Résumé

L'objectif de cet état de l'art est de présenter l'importance de la Segmentation Client dans les stratégies marketing, ainsi que d'explorer l'application de Data Mining sur ce domaine. Nous concentrons cette recherche sur le processus de création du modèle RFM et d'utilisation de l'algorithme K-means. Finalement, nous analysons les problèmes de l'algorithme K-means en proposant une version améliorée de cet algorithme.

Mots-clés : *segmentation client, data mining, algorithme K-means, modèle RFM*

Abstract

This state of art aims to present the importance of Customer Segmentation in the marketing strategies as well as to explore the application of Data Mining in this field. We focus this research on the process of the creation of the RFM model and the usage of the K-means algorithm. Finally, we analyze the problems of the K-means algorithm and we propose an improved version of this algorithm as well.

Keywords : *customer segmentation, data mining, K-means algorithm, RFM model*

Glossaire

CRM - Customer Relationship Management

ROI - Return on investment

TDM - Techniques de Data Mining

RFM - Recency, Frequency, Monetary

KDD - Knowledge Discovery in Databases

SCE - La Somme des Carrés de l'Erreur

1. Introduction

Les clients sont l'asset le plus important d'une entreprise. Il ne peut y avoir de perspectives commerciales sans clients satisfaits qui restent fidèles et développent une relation avec la société. C'est pourquoi une organisation doit planifier et appliquer une stratégie claire pour traiter les clients.

Les marketeurs organisent des campagnes de marketing direct pour communiquer des messages à leurs clients. Cela peut être mise en place par e-mail, télémarketing (téléphone), réseaux sociaux et d'autres canaux directs afin de stimuler l'acquisition de clients et l'achat de produits complémentaires (add-on products) et en même temps d'éviter le désabonnement.

Les marketeurs doivent avoir la capacité de comprendre les limites entre le marketing actif et le spam indésirable aux clients. En effet, si le modèle marketing est bien construit, il a la possibilité d'atteindre les clients ciblés et par conséquent d'augmenter le ROI.
En revanche, si le plan marketing n'identifie pas la bonne clientèle ou si cette dernière reçoit plusieurs campagnes dans une période courte, il est probable que le taux de désabonnement augmente.

Une des stratégies appliquée par les marketeurs afin de cibler leur audience est la segmentation client qui consiste à regrouper certains clients en fonction de caractéristiques semblables. Ces dernières années, de nombreuses recherches ont été menées pour améliorer la méthodologie du regroupement et pour bien définir les critères de la segmentation.
Grâce à ces travaux, des modèles statistiques ont été développés en utilisant les données clients afin de grouper la clientèle selon des critères différents, par exemple : comportementaux, socio-démographiques, la valeur de chaque client, les besoins etc.
Dans le monde du Big Data, un processus très utilisé de manipulation de données afin de trouver des patterns et des corrélations entre elles est celui de Data Mining. Plus précisément, on est concentré par la méthode de clustering qui signifie une analyse statistique utilisée pour organiser des données brutes en groupes homogènes

Il existe plusieurs approches afin d'effectuer de Data Mining, notamment cela dépend de l'objectif de l'entreprise. Un algorithme très utilisé afin de faire du clustering est celui du K-Means. Il permet de regrouper en K clusters distincts les observations du dataset de sorte que les données similaires se retrouvent dans le même cluster.

Dans ce contexte, quelle est l'importance de l'utilisation des méthodes de data mining pour la segmentation client ? Pourquoi l'algorithme K-means est-il très appliqué et quel est son problème associé ?
Comment pourrions-nous améliorer l'algorithme K-means ?

Dans une première partie, nous allons nous concentrer sur le concept de la Segmentation Client et son importance dans les stratégies marketing. Dans une deuxième partie nous allons présenter une définition de Data Mining ainsi que ses techniques descriptives et prédictives. Dans une troisième partie nous développerons plus en détail la relation entre la Segmentation Client et le Data Mining ainsi que l'application de ce dernier. Dans une dernière partie nous allons présenter une version améliorée de l'algorithme K-means.
Finalement, nous élaborons les conclusions de cet état de l'art en soulignant les enjeux de ce sujet et les perspectives des futurs travaux.

2. Méthodologie

L'objectif de cet état de l'art est d'effectuer une analyse approfondie de la littérature scientifique afin de rédiger une synthèse de documents concernant un sujet qui nous intéresse.

Dans un premier temps, nous avons sélectionné les domaines probables sur lesquels la recherche serait concentrée. Une liste préalable des domaines était : *la technologie cloud, les données, Big Data, les enjeux d'entreprise, le marketing.*
Ensuite nous avons exploré les sujets scientifiques que nous pouvions traiter afin de les aligner avec notre domaine d'intérêt.

Après avoir effectué une recherche sur les sujets traités et les articles scientifiques existants, nous avons décidé de choisir deux domaines parmi la première liste : *les données* et *le marketing*, deux domaines assez large. Il fallait donc formuler un sujet en choisissant une dimension spécifique pour chaque domaine. Concernant la structure du sujet, nous avons décidé de nous concentrer sur la formule "L'application de X à Y".
Après avoir réfléchi sur des thèmes potentiels, il était très intéressant d'analyser la relation entre les méthodes statistiques et les objectifs marketing.
Finalement, le sujet optimal était "L'application de Data Mining à la segmentation client".

En lisant des articles de recherche sur ce sujet, nous avons remarqué des différentes approches de Data Mining, notamment adaptées et alignées avec l'objectif marketing. Dans ces articles, il est mentionné très souvent l'application de l'algorithme K-means, mais en même temps d'autres papiers scientifiques présentaient des arguments pour lesquels cet algorithme n'est pas optimal.
Pour cette raison, nous avons trouvé intéressant de poser la question concernant la précision de l'algorithme K-means, mais aussi de trouver des méthodes d'amélioration.

Le canal principal de recherche était Google Scholar et le processus de recherche a été effectué en anglais. Nous avons utilisé des mots clés tels que *client segmentation*, *data mining*, *k-means algorithm*, *customer relationship management*. Etant donnée que le domaine de Data Mining et celui de la Segmentation Client sont assez larges, il était nécessaire d'effectuer une recherche

plus spécifique. Ainsi, des expressions clés sont utilisés tels que : *data mining in customer segmentation*, *the importance of customer segmentation*, *an improved k-means algorithm*, *the techniques of data mining*.

Finalement, 8 articles de recherche et un livre ont été retenus comme références essentielles de cet état de l'art.

3. CRM - Customer Relationship Management

Il existe un grand nombre d'études et de définitions concernant le CRM. Selon [1] le CRM signifie une stratégie globale et un processus d'acquisition, de fidélisation et de partenariat avec des clients sélectionnés pour créer une valeur supérieure pour l'entreprise et le client. Cela implique l'intégration du marketing, des ventes, du service client et des fonctions de la chaîne d'approvisionnement de l'organisation pour atteindre une plus grande efficacité dans la création de valeur client.

Selon [2] le CRM est fondé sur 3 grands composants :

- La technologie fait référence aux capacités informatiques qui permettent à une entreprise de collecter, d'organiser, de sauvegarder et d'utiliser des données sur ses clients.
- Les clients et les employés. Le CRM est construit autour des clients afin de générer des relations bénéfiques en acquérant des informations sur différents aspects des clients. Néanmoins, un engagement total du personnel et de la direction de l'organisation est essentiel pour une mise en œuvre efficace du CRM afin de mieux servir les clients et de satisfaire leurs besoins.
- Le processus métier. Le succès du CRM nécessite un changement des processus métier vers une approche centrée sur le client.

[3] a divisé le CRM en quatre dimensions différentes, illustrées sous forme d'un modèle itératif (voir figure 1) :

1. Identification des clients (Customer identification)
2. Attraction des clients (Customer attraction)
3. Fidélisation des clients (Customer retention)
4. Développement des clients (Customer development)

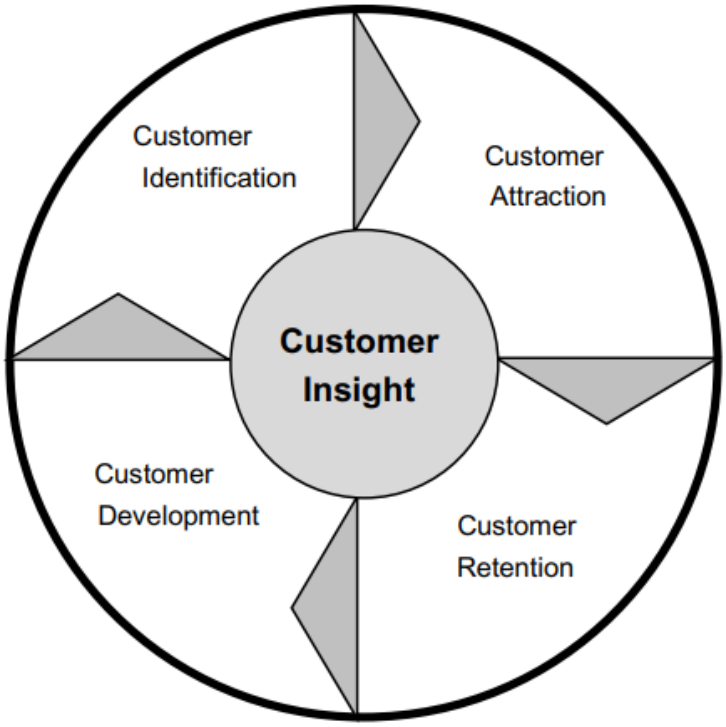

Figure 1. Le concept du Customer Management [3]

[4] est basé sur le modèle de [3] pour préciser la notion de segmentation client. Selon [4], **l'identification des clients** est extrêmement importante car l'identification des propriétés des clients aide les entreprises à sélectionner les stratégies appropriées. **La segmentation des clients** est l'une de ses parties les plus importantes et comprend la division de l'ensemble des clients en plusieurs segments de valeur.

4. Segmentation Client

4.1. Définition

La segmentation client n'est pas une notion récente. Ce concept est attribué à Wendell R Smith [5] où il a préconisé d'appliquer la segmentation de la clientèle ainsi que la différenciation des produits. Depuis lors, un grand nombre de recherches ont été menées. Danaher dans sa recherche [6] examine les méthodes de la segmentation où il analyse le modèle de l'industrie aérienne et celui de la télécommunication. [7] définit la segmentation de la clientèle comme suit : "La segmentation de la clientèle est le processus de division des clients en sous-groupes distincts, significatifs et homogènes en fonction de divers attributs et caractéristiques." Elle peut être appliquée afin d'optimiser l'allocation des ressources, de développer des stratégies de marketing appropriées et de fournir des services appropriés pour chaque groupe de clients [8].
D'ailleurs, selon [9] la segmentation de la clientèle se concentre sur les critères comportementaux de la segmentation du marché car elle se concentre sur les caractéristiques comportementales des clients.

4.2. Les types de segmentation client

Il existe différents critères qui peuvent être pris en compte afin de construire des stratégies marketing sous plusieurs types de segmentation.
Selon [7] les types de segmentation suivants sont les plus utilisés :

- **Basé sur la valeur** : dans la segmentation basée sur la valeur, les clients sont regroupés en fonction de leur valeur. C'est l'un des types de segmentation les plus importants car il peut être utilisé pour identifier les clients les plus précieux et pour suivre la valeur et les changements de valeur au fil du temps. Il est également utilisé pour différencier les stratégies de prestation de services et pour optimiser l'allocation des ressources dans les initiatives de marketing.

Dans [10], les auteurs insistent sur le fait que cette segmentation n'est pas d'une tâche ponctuelle. Il est vital pour l'organisation d'être en mesure de suivre les changements de valeur dans le temps.

- **Comportemental** : Il s'agit d'un type de segmentation très efficace et utile. Il est largement utilisé car il présente des difficultés minimales en termes de disponibilité des données. Les données requises comprennent les données de propriété et d'utilisation des produits qui sont généralement stockées et disponibles dans les bases de données de l'organisation. Les clients sont divisés en fonction de leurs modèles de comportement et d'utilisation des produits identifiés. Ce type de segmentation est généralement utilisé pour développer des stratégies d'offre de produits personnalisées mais aussi, pour le développement de nouveaux produits et la conception de programmes de fidélité.

- **Basé sur la propension :** dans la segmentation basée sur la propension, les clients sont regroupés en fonction de scores de propension, tels que les scores de désabonnement, les scores de vente croisée, etc., qui sont estimés par des modèles de classification (propension) respectifs. Les scores de propension peuvent également être combinés avec d'autres schémas de segmentation pour mieux cibler les actions marketing.

- **Basé sur la fidélité :** la segmentation de la fidélité implique l'étude du statut de fidélité des clients et l'identification des segments basés sur la fidélité tels que : les fidèles et les migrants. Les segments peuvent être associés à des comportements d'utilisation spécifiques et à des attributs de base de données client. Pour y parvenir, une organisation peut commencer par une étude de marché pour révéler les segments de fidélité, puis construire un modèle de classification avec le champ des segments de fidélité comme cible. Ainsi, il pourra identifier les comportements associés à chaque segment de fidélité et utiliser les règles de classification pertinentes pour extrapoler les résultats de la segmentation de fidélité à l'ensemble de la clientèle.
  [7] a donné une représentation schématique des segments basés sur la fidélité d'une entreprise téléphonique (figure 2) où les fidèles et les migrants sont segmentés en fonction du comportement de la clientèle.

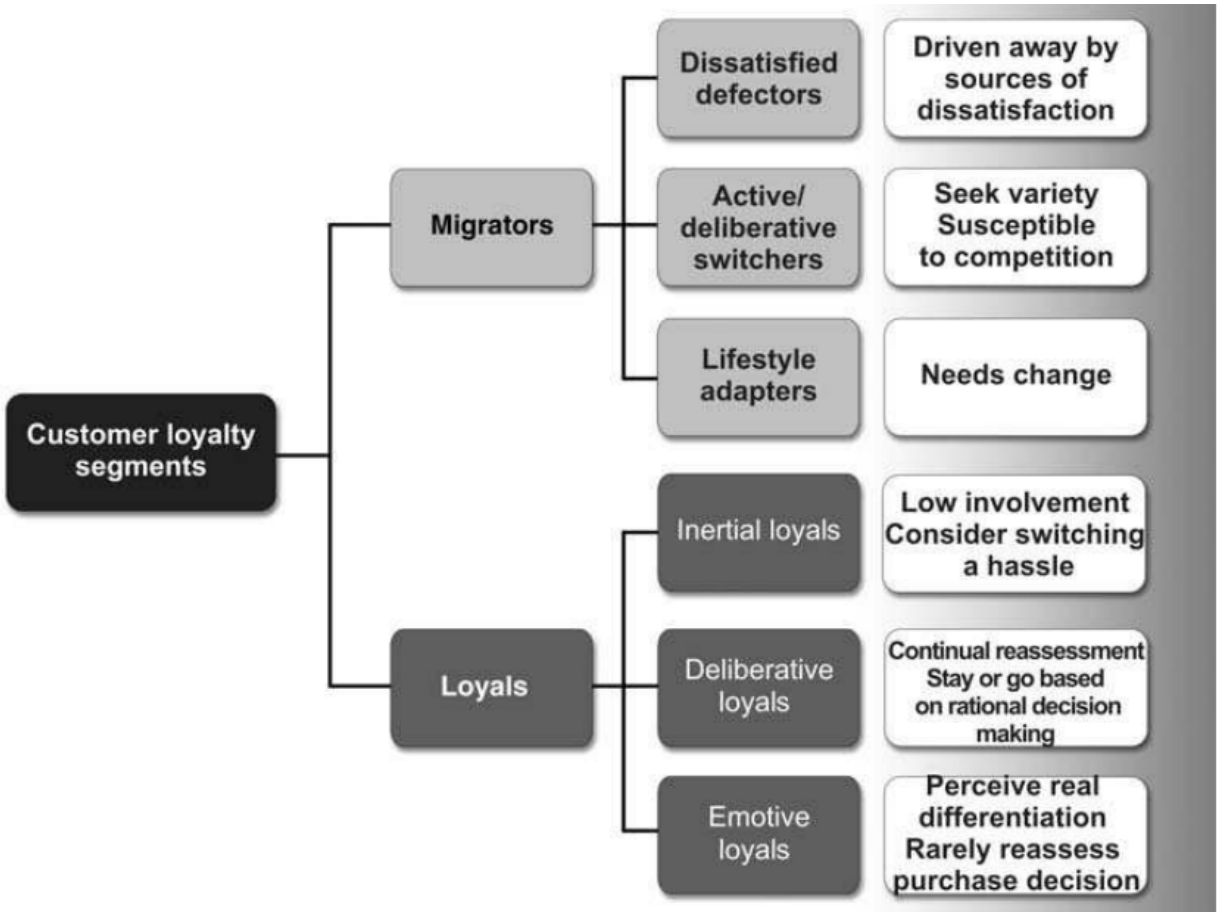

Figure 2. Les segments basés sur la fidélité [7]

- **Socio-démographique et étape de vie :** ce type révèle différents groupes de clients en fonction d'informations socio-démographiques et/ou d'étapes de vie telles que l'âge, le revenu, l'état civil. Ce type de segmentation est approprié pour promouvoir des produits spécifiques basées sur des spécificités socio-démographiques.

- **Besoins/attitude :** ce type de segmentation est généralement basé sur des données d'études de marché et identifie les segments de la clientèle en fonction de leurs besoins, désirs, attitudes, préférences et perceptions concernant les services et produits de l'entreprise. Il peut être utilisé pour soutenir le développement de nouveaux produits et pour déterminer l'image de marque et les principales caractéristiques du produit à communiquer.

4.3. L'importance de la segmentation client dans la stratégie marketing

La segmentation de la clientèle est une approche appliquée dans plusieurs industries qui se concentrent sur la relation client.

4.4. Le secteur bancaire

La concurrence croissante dans le système bancaire et dans les institutions financières non bancaires nécessite l'utilisation de stratégies marketing et d'approche client [11]. Selon cette étude, récemment une des priorités principales et une condition de réussite des banques au détail est l'amélioration de la segmentation de la clientèle. Elle permet aux institutions financières de réguler l'offre de services en fonction des utilisateurs actuels et potentiels du marché et de développer des stratégies de marché à long terme. La segmentation du marché bancaire est aujourd'hui un facteur clé pour le développement d'une entreprise prospère [11].

4.4.1. Le processus de la segmentation

Considérant ce qui précède, un outil commun pour améliorer la compétitivité est la conception d'une gamme spéciale de produits et de services ciblant les clients fidèles ou leur offrant des remises spéciales pour les produits existants. Cela se transcrit à travers les « programmes de fidélité ».

Selon la définition de [12], un programme de fidélité est un programme de marketing conçu pour fidéliser la clientèle en offrant des incitations aux clients rentables. Un programme de fidélité repose souvent sur plusieurs propositions, telles que les suivantes :

1. Les clients peuvent vouloir des relations avec les produits qu'ils achètent.
2. Une partie de ces clients a tendance à être fidèle.
3. Ils forment un groupe rentable.
4. Il est possible de renforcer la fidélité de ces clients grâce au programme de fidélité.

Pour augmenter le nombre de clients fidèles, les banques ont commencé à expérimenter diverses méthodes et outils.

Dans sa recherche, [11] a pris en étude une base de données de 100 emprunteurs d'une agence bancaire commerciale ayant contracté des prêts à la consommation garantis. Les clients sont définis comme fidèles en fonction de leur historique de crédit (ils ont moins de 3 paiements manqués pour la dernière année). Le but de cette analyse est de segmenter les emprunteurs en 3 groupes en utilisant les 3 variables suivantes :

- Le Montant du prêt (en euros)
- Depuis combien de temps l'emprunteur est client de la Banque (en mois)
- Un état représentant le nombre de paiements manqués par client au cours des 12 derniers mois.
    - 0 -> 0 paiement manqué
    - 1 -> 1 paiement manqué
    - 2 -> 2 paiements manqués

Après avoir appliqué des modèles statistiques qui seront développés dans cet état de l'art, [11] a réussi catégoriser les emprunteurs de la base de données initiale, dans 3 groupes suivants :

1. Clients Platinum
2. Clients Gold
3. Clients Silver

En ayant une information précise sur chaque segment, les managers des programmes de fidélité peuvent appliquer des stratégies de marketing tel que l'encouragement de ces clients fidèles.

4.5. Le secteur retail

L'industrie retail est un autre secteur où la segmentation de la clientèle est bien appliquée.

[13] a effectué une étude à partir d'un ensemble de données des comptes de cartes de fidélité des clients de la base de données d'une entreprise retail de sport. Dans leur étude, ils ont utilisé des données recueillies par une chaîne de magasins retail. Cette dernière est l'une des plus grandes de Turquie dans le commerce de détail de sport.

Comme toutes les autres entreprises de vente retail de sport, la société propose des produits tels que des chaussures, des chemises, des pulls, des accessoires et des équipements sportifs. Les gestionnaires avaient décidé de créer un système de carte de fidélité client dans le but de les segmenter. Le programme de cartes de fidélité se composait de trois niveaux de cartes : bronze, or et prime.

4.5.1. Le processus de la segmentation

Le but de cette segmentation est de grouper la base de données d'acheteurs récents en trois catégories afin de créer un système de fidélité client. Les clients membres du programme de fidélité ont été groupés à partir des points qu'ils gagnent en fonction de leurs dépenses au cours d'une année civile.

Le système actuel de fidélité est comme suivant : Les clients qui ont une carte Bronze sont les membres qui ont dépensé moins de 2000 livres turques (TL) (≈ 520 $) en un an. Les membres de la carte Or sont les clients qui ont dépensé entre 2 000 et 4 000 TL (≈ 520 $ à 1 040 $). Les clients qui ont dépensé plus de 4 000 TL (≈ 1 040 $) méritent d'avoir une carte Premium. Cela indique que leur segmentation n'est basée que sur les dépenses totales de la clientèle. Selon la segmentation actuelle de l'entreprise, sur 700033 clients, 694647 ont une carte bronze, 4469 ont une carte or et 916 ont une carte premium.

Après avoir appliqué des modèles statistiques lesquels seront développés dans cet état de l’art, [13] a obtenu les résultats suivants.
La démographie de la population :

- 62,04% d'hommes et 37,96% de femmes.
- L'âge des clients variait de 16 à 74 ans.
- La majorité des consommateurs (85,5%) avait moins de 30 ans.

L'historique d'achats des consommateurs :

- La dépense moyenne des clients est de 336 TL (≈ 87 $).
- Ils achètent environ 2 (1,93) fois en un an.

- La carte de crédit est l'outil de paiement préféré avec 59,9 %, suivie par les espèces (25,13 %) et les cartes-cadeaux (14,65 %).
- La majorité des clients (86,8%) ont préféré faire leurs achats dans des magasins plutôt que des achats en ligne.

L'étude a considéré des facteurs comme les dépenses des clients, la fréquence de leurs achats et la date de leur dernier achat. Finalement, [13] a réussi à diviser la base de données client dans quatres segments différents. Les segments suivants sont dans un ordre croissant en termes de la valeur qu'ils portent pour l'entreprise en considérant les trois facteurs.

- Réguliers - 644081 clients
- Fidèles - 514 clients
- Stars - 97 clients
- Avancés - 55340 clients

Ils ont constaté que dans le premier segment Réguliers, on y trouve les clients ayant eu des indicateurs en dessous du niveau moyen. Les managers peuvent donc réfléchir sur comment traiter cette catégorie. Nous pouvons bien remarquer la différence entre le système actuel de fidélité et celui après avoir appliqué les méthodes statistiques de la segmentation client.

4.6. Le secteur e-commerce

[14] présente l'application de la segmentation client dans le secteur de l'e-commerce, plus spécifiquement des sites web de cashback. Etant donné que le concept "site web cashback" est assez récent, il n'existe pas beaucoup de papiers de recherche, ni de définition scientifique.
Selon [14] les sites web de cashback sont basés sur un type spécifique de marketing d'affiliation, qui est une pratique de marketing basée sur le web dans laquelle une entreprise récompense un ou plusieurs affiliés pour chaque visiteur ou client généré par les efforts de marketing de l'affilié.

Les sites Web affiliés servent d'outil pour attirer les clients, de la même manière que les annonces des moteurs de recherche. Les consommateurs accèdent aux sites affiliés (sites de cashback) au lieu du site Web des retail. Les offres sur ces sites sont négociées à l'avance avec les retailers et

publiées sur le site Web, généralement avec le cashback qui sera livré lié. La société de cashback reçoit une commission sur chaque transaction effectuée et dépose les paiements en espèces directement sur les comptes bancaires des consommateurs [15].

Selon [14], la segmentation est basée sur deux critères : l'activité commerciale des clients et leur rôle au sein du réseau social du site. Cette étude montre comment le rôle du client au sein du réseau social du site de cashback détermine le comportement et l'activité commerciale du client sur le site.

4.6.1. Le processus de la segmentation

La segmentation présentée décrit le parcours client en termes de rentabilité client et d'ancienneté. [14] ont utilisé les données de l'un des plus grands sites de cashback d'Europe continentale. Ce site Web est présent dans 14 pays, compte plus de deux millions de clients et réalise un chiffre d'affaires annuel de plus de 20 millions d'euros. Les observations ont été recueillies de janvier à mars 2015. Des données sur toutes les transactions des clients et leurs rôles au sein du réseau social ont été collectées. Pour éviter les biais d'échantillonnage, les nouveaux clients qui se sont inscrits sur le site durant cette période ont été exclus car ces clients n'avaient pas eu suffisamment de temps pour développer leurs réseaux sociaux.

Les clients ont été regroupés en huit clusters en fonction de leur activité commerciale et de leur rôle au sein du réseau social du site de cashback.
Finalement, les clients du site sont groupés dans huit segments différents.

- Groupe 1 : Acheteurs référés immatures.
- Groupe 2 : Les gros utilisateurs qui investissent du temps sur le site (faible rentabilité).
- Groupe 3 : Les gros utilisateurs de toutes sortes de transactions avec une faible sensibilité aux paiements (rentabilité élevée).
- Groupe 4 : Utilisateurs axés sur la rentabilité et la commodité (rentabilité moyenne à élevée).
- Groupe 5 : Acheteurs de proximité avec potentiel.

- Groupe 6 : Acheteurs référés en cours de développement.
- Groupe 7 : Acheteurs de proximité engagés.
- Groupe 8 : Acheteurs référés engagés.

Cette analyse a plusieurs implications pour les praticiens, non seulement dans les sites de cashback, mais aussi dans le marketing d'affiliation. Les résultats montrent aux managers comment traiter différents clients aux caractéristiques différentes pour renforcer leur fidélité et leur contribution à une marque en plein développement. Les résultats peuvent générer des rendements particulièrement élevés pour les affiliés dans un environnement de plus en plus concurrentiel.

5. Data Mining

5.1. Définition

Selon [16], le Data Mining est la recherche des relations et des modèles qui existent dans des grandes bases de données qui sont noyées à cause de l'immensité des données.
Il consiste à découvrir des modèles potentiellement utiles et à appliquer des algorithmes pour l'extraction de ces informations [17].

[20] donne une définition plus large. Le Data Mining, qu'on rapproche souvent du KDD "découverte de connaissances dans les bases de données" (KDD - Knowledge Discovery in Databases), fait référence à l'extraction non triviale d'informations implicites, auparavant inconnues et potentiellement utiles à partir de données dans des bases de données. Alors que le Data Mining et le KDD sont souvent traités comme des synonymes, le Data Mining fait en fait partie du processus de découverte de connaissances. La figure suivante (figure 3) montre l'exploration de données comme une étape d'un processus itératif de KDD.

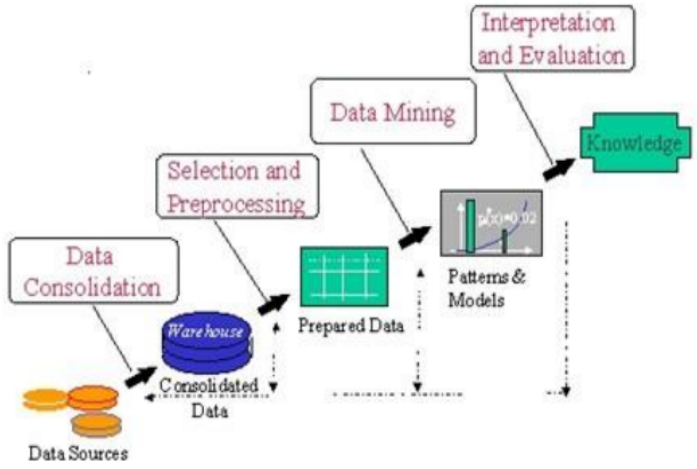

Figure 3. Le Data Mining est au cœur du processus de découverte des connaissances (KDD) [20]

5.2. Les techniques de Data Mining

Les techniques de Data Mining sont utilisées pour extraire des patterns à partir de données massives et le résultat peut être divisé en différents types en fonction de son objectif et de ses exigences [18]. En se basant sur leurs fonctions, les techniques de Data Mining sont divisées en techniques descriptives et en techniques prédictives (voir figure 4).

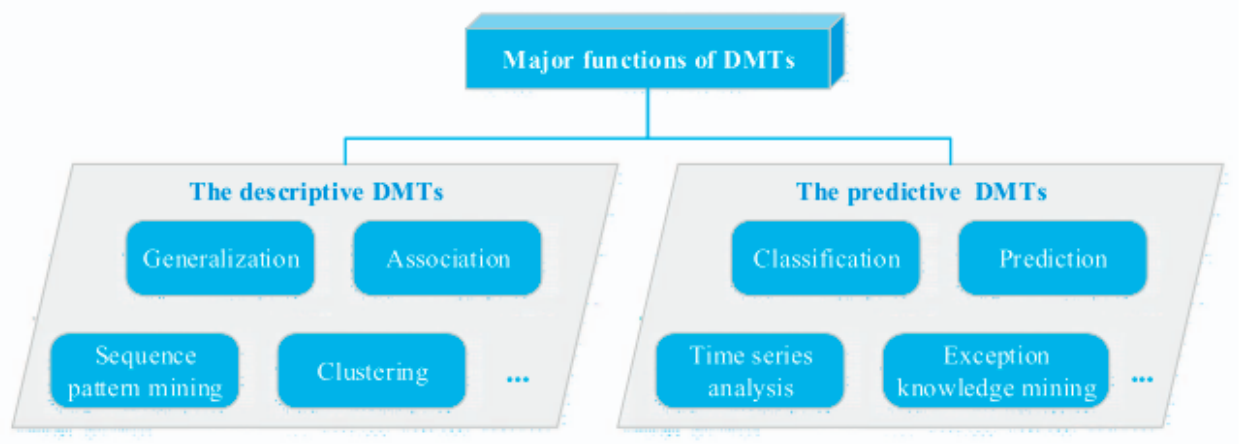

Figure 4. La classification des techniques de Data Mining [18]

5.2.1. Les techniques descriptives de Data Mining

Les fonctions descriptives des TDM (techniques de data mining) visent principalement à explorer les règles, caractéristiques et relations potentielles ou récessives (telles que la dépendance, la similarité, etc.) qui existent dans les données, telles que la généralisation, l'association, l'exploration de modèles de séquences, le regroupement, etc [18].

La généralisation

Ce concept fait référence à la généralisation des connaissances extraites d'une base de données. Son objectif est de trouver un lien universel entre les données détaillées qui sont déjà stockées. Le résultat peut servir de base à l'application d'autres types de technique telles que la classification et la prédiction.

Les méthodes de généralisation existantes comprennent principalement la méthode de description, l'analyse de données multidimensionnelles et la technologie de stratification [18].

L'association

L'association implique la dépendance ou l'association entre plusieurs événements. Le but de l'association est de trouver l'interdépendance entre des caractéristiques ou des données à partir d'un grand nombre de données.

Basée sur différentes règles d'association, elle peut être divisée en association simple, association temporelle, association causale, etc. [18]

Mining sequential pattern

Mining sequential pattern est une TDM utilisée pour découvrir des sous-séquences temporelles à haute fréquence ou d'autres séquences liées à partir de la base de données. Autrement dit, selon [19] Mining sequential pattern est une technique d'exploration de données utilisée pour identifier des modèles d'événements ordonnés dans une base de données. Son application provient à l'origine de l'industrie de la vente retail ou elle pouvait être utilisée pour savoir si un client allait acheter la suite d'un livre lors de son achat dans un certain laps de temps [19].

Il est semblable à l'association. Cependant, son objet est une séquence avec un certain ordre et il est prédictif [18].

Clustering

Le clustering divise un groupe d'individus en plusieurs catégories en fonction de leurs similitudes et où les différences entre ces individus d'une même catégorie doivent être aussi petites que possible [18].

Les méthodes les plus représentatives du clustering sont basées sur la mesure de distance géométrique. Les techniques de clustering peuvent analyser des modèles de données d'entrée complexes et suggérer des solutions qui ne seraient pas évidentes autrement. Ils révèlent des typologies de clients, permettant des stratégies marketing très efficaces [7].

Les applications du clustering se répartissent entre le marketing, l'urbanisme, la psychologie, la recherche archéologique, la recherche sur les tremblements de terre, la recherche

météorologique, etc. L'algorithme K-means de cette technique sera développé plus en détail dans cet état de l'art.

5.2.2. Les techniques prédictives de Data Mining

Les techniques prédictives des TDM sont généralement appliquées afin d'analyser les tendances pertinentes des données ou les lois pertinentes afin de prédire l'état futur. Il comprend la classification, la prédiction, l'analyse des séries temporelles, les exceptions, etc. Ce type de TDM présente la tendance à partir de données existantes dans le but de prédire leurs classifications futures ou la valeur continue en fonction des tendances inférées des données [18].

La classification

Le but de la classification est de construire une fonction ou un modèle selon les caractéristiques de l'ensemble des données, et ensuite de catégoriser chaque objet dans une classe d'objets connue. La classification a un large éventail d'applications, telles que le diagnostic médical, l'évaluation de solvabilité pour les crédits, la reconnaissance des modèles d'images, le positionnement sur le marché cible, la détection des défauts, l'analyse de l'efficacité, le traitement graphique, l'analyse des fraudes à l'assurance, etc [18].

La prédiction

La prédiction consiste à exploiter les connaissances générées par les données historiques et actuelles afin de déduire les tendances futures des données. Bien que la classification soit utilisée pour prédire les classes, les analystes souhaitent souvent prédire certaines valeurs de données manquantes ou inconnues. En d'autres termes, le résultat de prédiction souhaité correspond aux données numériques. [18]

La prédiction est utilisée dans de nombreux domaines, tels que l'identification d'objets dans de grandes bases de données d'images, l'évaluation de solvabilité pour les crédits, le diagnostic médical, la prédiction de performance, le marketing, etc.

L'analyse des séries temporelles

L'analyse de séries temporelles est l'extraction d'informations et de connaissances à partir d'un grand nombre de données de séries temporelles avec un ou plusieurs attributs temporels potentiellement utiles pour la prédiction à court, moyen et long terme. La série temporelle est une forme spéciale de données où les valeurs passées de la séquence affectent les valeurs futures. L'analyse des séries temporelles est principalement utilisée pour résoudre deux types de problèmes. L'une consiste à résumer la séquence ou la tendance des données, comme l'étude du comportement d'achat des clients des supermarchés, la prévision des stocks, les transactions futures, la prévision des enregistrements de séquence de clics de page, etc. L'autre consiste à surveiller le changement périodique des données [18].

Exception knowledge mining

Exception knowledge mining concerne un cas particulier extrême trouvé dans les données sources qui se distingue des autres données, ce qui révèle les réponses concernant les objets qui ne suivent pas la loi normale.

L'exception knowledge mining y compris l'analyse des valeurs aberrantes, l'analyse des anomalies de séquence, la découverte de règles spécifiques, etc., est un processus qui permet de découvrir des comportements différents de ceux attendus. Il peut être combiné avec d'autres TDM pour acquérir davantage de connaissances sur les exceptions tout en creusant des connaissances communes, par exemple, les cas d'anomalies dans la classification, les cas particuliers sans respecter les règles générales, les données aberrantes de regroupement de données, etc. Il est d'une grande valeur dans certains domaines, comme la fraude à l'assurance et à la carte de crédit, l'approbation de prêt, l'analyse médicale, la sécurité du réseau, la détection d'intrusion, l'exploration de texte dans la découverte de nouveaux thèmes, etc. [18]

6. Le Data Mining pour la Segmentation Client

Les articles [4][13][21][22][23][28] ont des approches similaires au niveau de la méthodologie suivie afin d'effectuer le processus de Data Mining.

Ce processus est présenté sur la figure 5.

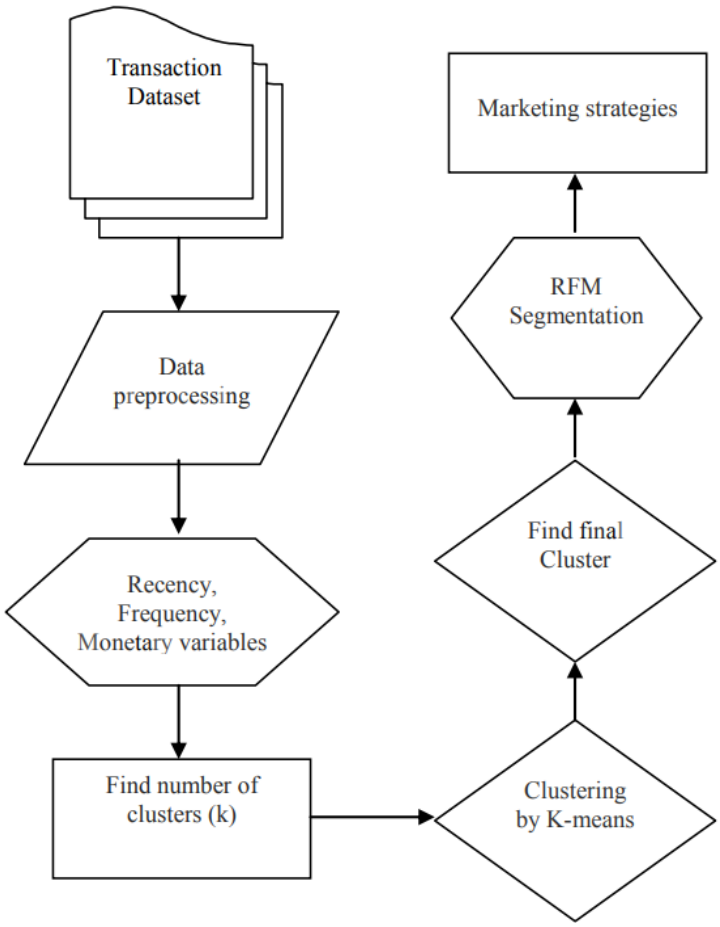

Figure 5. Processus de segmentation client basé sur le modèle RFM et les techniques de clustering [23]

6.1. Le modèle RFM

L'analyse de variables de récence, de fréquence et monétaire (RFM) est une technique puissante et reconnue dans le marketing de bases de données. Elle est largement utilisée pour classer les clients en fonction de leurs antécédents d'achat [21] ainsi qu'analyser la valeur client [22]. L'analyse RFM trouve une utilisation dans un large éventail d'applications impliquant un grand nombre de clients tels que l'achat en ligne, la vente au détail, etc. Cette méthode regroupe les clients en fonction de trois dimensions, récence (R), fréquence (F) et monétaire (M).
[22] a appliqué le RFM afin de segmenter et d'analyser le comportement du client en fonction des données comportementales des transactions, avec l'objectif de déterminer les clients clés ou stratégiques. [23] a utilisé le modèle RFM pour traiter des données transactionnelles d'une entreprise. Une base de données de 82,648 transactions a été analysée et après avoir appliqué le RFM, le résultat indiquait qu'elles avaient été réalisées par 102 clients. Ce résultat a été utilisé pour appliquer par la suite des algorithmes de clustering.

La récence - Quelle est la date du dernier achat du client ?
La valeur de récence est le nombre de jours qu'un client prend entre deux achats. Une valeur de récence plus faible implique que le client visite l'entreprise à plusieurs reprises sur une courte période. De même, une valeur plus élevée implique que le client est moins susceptible de visiter l'entreprise. [21][22]

Fréquence – Combien de fois le client a-t-il acheté ?
La fréquence est définie comme le nombre d'achats (transactions [22]) qu'un client effectue au cours d'une période spécifique. Plus la valeur de la fréquence est élevée, plus les clients de l'entreprise sont fidèles [21][22].

Monétaire – Combien d'argent le client a-t-il dépensé ?
La valeur monétaire est définie comme la somme d'argent dépensée par le client pendant une certaine période. Plus le montant d'argent dépensé est élevé, plus ils rapportent à l'entreprise. [21][22]

6.1.1. L'application du modèle RFM

Le but du modèle RFM est de créer un système de notation [21][22][23], afin que le résultat puisse être utilisable ensuite dans des algorithmes de clustering.

La base initiale de données de [23] est celle des transactions effectuées dans l'entreprise Nine Reload. Un fragment de ces données est illustré sur la figure 6.

Id	Date	Name	Hp	Id Member	Product	To	Status	Suplier	Price	Repeat	Des	Sn	Profit	beginning balance	ending balance	Com	Ip	price package	Type
22370	31-01-201	SAHAL CELL	6.2857E+12	SR00038	I5	8.5703E+10	sukses	SEMBILAN R	5,675	0	22-07-0	7.61E+17	225	155,671	149,996	207	127.0.0.	HARGA MA	Saldo
22369	31-01-201	CAHAYA CELL	6.2857E+12	SR00110	PLN20	4.5002E+10	sukses	SEMBILAN R	20,550	0	21:43:3	1224-794	675	30,905	10,355	207	127.0.0.	HARGA MA	Saldo
22368	31-01-201	TARI	6.2877E+12	SR00134	X100	8.7879E+10	sukses	SEMBILAN R	100,150	0	21:37:0	6.65E+13	2,400	338,550	238,400	207	127.0.0.	HARGA MA	Saldo
22367	31-01-201	IBU CELL	6.2877E+12	SR00008	X5	8.1809E+10	sukses	SEMBILAN R	5,700	0	21:33:1	1.7E+13	200	193,500	187,800	207	127.0.0.	HARGA MA	Saldo
22366	31-01-201	ADI CELL	6.2856E+12	SR00017	I10	8.5648E+10	sukses	SEMBILAN R	10,675	0	21:17:4	7.61E+17	225	304,160	293,485	207	127.0.0.	HARGA MA	Saldo
22365	31-01-201	ADI CELL	6.2856E+12	SR00017	SN20	8.2135E+10	sukses	SEMBILAN R	20,300	0	20:59:0	4.1E+13	500	324,460	304,160	207	127.0.0.	HARGA MA	Saldo
22364	31-01-201	ADI CELL	6.2856E+12	SR00017	SN50	8.2135E+10	sukses	SEMBILAN R	49,800	0	20:58:1	4.1E+13	1,650	374,260	324,460	207	127.0.0.	HARGA MA	Saldo
22363	31-01-201	HUYA CELL	6.2859E+12	SR00097	I5	8.5889E+10	sukses	SEMBILAN R	5,775	0	20:51:1	7.62E+17	325	232,159	226,384	207	127.0.0.	HARGA MA	Saldo
22362	31-01-201	HUYA CELL	6.2859E+12	SR00097	IS10	8.5641E+10	sukses	SEMBILAN R	10,800	0	20:50:3	7.61E+17	300	242,959	232,159	207	127.0.0.	HARGA MA	Saldo
22357	31-01-201	LIDA CELL	6.2856E+12	SR00039	I10	8.5871E+10	sukses	SEMBILAN R	10,775	0	20:37:3	7.61E+17	325	102,074	91,299	207	127.0.0.	HARGA MA	Saldo
22356	31-01-201	LIDA CELL	6.2856E+12	SR00039	SN10	8.5327E+10	sukses	SEMBILAN R	10,700	0	20:36:3	4.1E+13	475	112,774	102,074	207	127.0.0.	HARGA MA	Saldo
22353	31-01-201	INA CELL	6.2838E+12	SR00123	X10	8.7803E+10	sukses	SEMBILAN R	10,800	0	20:26:3	1.7E+13	300	27,870	17,070	207	127.0.0.	HARGA MA	Saldo
22352	31-01-201	FAIZAL CELL	262478663	SR00098	I5	8.5718E+10	sukses	SEMBILAN R	5,675	0	20:12:1	7.62E+17	225	50,825	45,150	207	127.0.0.	HARGA MA	Saldo
22351	31-01-201	AJENG CELL	241838947	SR00109	I5	085726371	gagal	SEMBILAN R	5,825	0	19:55:0	7.62E+17	375	57,447	57,447	207	127.0.0.	HARGA MA	Saldo
22350	31-01-201	HUYA CELL	6.2859E+12	SR00097	IS5	8.5866E+10	sukses	SEMBILAN R	5,800	0	19:54:1	7.62E+17	300	48,759	42,959	207	127.0.0.	HARGA MA	Saldo
22349	31-01-201	ANES CELL	6.2877E+12	SR00107	T10	8.9538E+11	sukses	SEMBILAN R	10,300	0	19-23-2	1.31E+17	430	1,300,220	1,289,920	207	127.0.0.	HARGA MA	Saldo
22348	31-01-201	HILYA CELL	6.2852E+12	SR00015	SN10	8.2345E+10	sukses	SEMBILAN R	10,600	0	19-21-0	4.1E+13	375	159,575	148,975	207	127.0.0.	HARGA MA	Saldo
22347	31-01-201	TASY CELL	6.2857E+12	SR00077	SN10	8.2227E+10	sukses	SEMBILAN R	10,600	0	19:19:4	4.1E+13	375	74,827	64,227	207	127.0.0.	HARGA MA	Saldo
22346	31-01-201	INA CELL	6.2838E+12	SR00123	I10	8.5817E+10	sukses	SEMBILAN R	10,775	0	19:15:4	7.61E+17	325	38,645	27,870	207	127.0.0.	HARGA MA	Saldo
22345	31-01-201	AJENG CELL	241838947	SR00109	I5	8.5726E+10	sukses	SEMBILAN R	5,825	0	19:15:3	7.62E+17	375	63,272	57,447	207	127.0.0.	HARGA MA	Saldo

Figure 6. Les données de transaction [23]

La première étape est la préparation de données. Tout d'abord, les variables les plus pertinentes pour le modèle RFM ont été choisies. Les variables nécessaires, ainsi que des informations complémentaires se trouvent sur la table 1.

Field	**Information**
Champ	**Information**
Agent name	Used to specify the customer code.
Le nom de l'agent	Utilisé pour spécifier le code client.

Date Date	The date of the customer's purchase transaction is used to model Recency and Frequency. - Recency, within a year when the last customer made a transaction with Nine Reload. - Frequency is the number of transactions conducted by the customer within a period of one year. La date de la transaction d'achat du client est utilisée pour modéliser la récence et la fréquence. - La récence signifie la dernière transaction d'un client avec Nine Reload sur une période d'un an. - La fréquence est le nombre de transactions effectuées par le client sur une période d'un an.
Price Prix	This variable is used to model the Monetary attribute, by summing up all customers transactions in one year. Cette variable est utilisée pour modéliser l'attribut Monétaire, en additionnant toutes les transactions des clients sur une année.

Table 1. Les variables utilisées. [23]

Le total de 82,648 transactions est ensuite filtré grâce aux variables du RFM mentionnées ci-dessus.

Le résultat obtenu de 102 clients est présenté sur la table 2.

Agent Code	**R**	**F**	**M**
C001	31-12-2017	2035	Rp 22,909,504.00 ≃ 1500.46 €
C002	18-06-2017	339	Rp 5,878,306.00 ≃ 385.00 €
C003	04-11-2017	352	Rp 4,525,250.00 ≃ 296,38 €
C004	31-12-2017	36	Rp 526,250.00 ≃ 34,47 €
....	....	....	....
C102	25-01-2017	28	Rp 231,375.00 ≃ 15,15 €

Table 2. Le résultat obtenu après avoir appliqué les variables RFM [23]

Après avoir obtenu le résultat conformément aux variables RFM, il est essentiel de construire un système de notation afin que les données soient au même format numérique. Pourtant, ce système n'est pas standard, il peut varier selon l'objectif et le niveau de complexité du modèle.

Dans [23], il est construit un système de notation comme suivant :

- Ils définissent un intervalle de poids entre 1 (min) et 5 (max).
- Ils précisent pour chaque variable RFM, les critères respectifs en créant 5 catégories de valeurs client différents.
- Chaque catégorie est associée avec des poids spécifiques selon la valeur que chacun porte, 1 pour la catégorie de la valeur d'un client faible jusqu'à 5 pour celle d'une grande valeur.

Cela veut dire que le client avec la plus grande valeur aura des poids (5, 5, 5) (R, F, M) et celui avec la plus petite valeur (1, 1, 1) Les catégories avec les poids respectifs sont présentées sur la table 3.

Poids	**R**		**F**		**M**	
5	Shortest	< 1 Month	Highest	> 15000	So Many	> 300 Million
4	Short	1 - 3 Month	High	8000 - 15000	Many	150 - 200 Million
3	Regular	3 - 5 Month	Regular	5000 - 8000	Normal	100 - 150 Million
2	Long	5 - 8 Month	Low	2000 - 5000	Few	50 - 100 Million
1	Longest	> 8 Month	Lower	< 2000	Fewer	< 50 Million

Table 3. La table de décision [23]

La table 4 présente un exemple des résultats, après avoir appliqué le système des poids.
Nous remarquons que le client avec la plus grande valeur a les poids (5, 2, 1) et le plus bas a les poids (1, 1, 1).

Agent Code	R	F	M
C001	5	2	1
C002	1	1	1
C003	1	1	1
C004	5	1	1
....	....	....	....
C102	1	1	1

Table 4. Le résultat du système de notation [23]

Une fois que les données sont transformées en valeurs numériques, elles sont pertinentes pour être utilisées par l'algorithme K-means. Ce dernier sera traité plus en détail dans cet état de l'art.

Dans [22], une autre approche est appliquée par les auteurs afin de créer leur système de notation. Ils ont traité une base de données de 71 161 transactions sur une période de 7 mois et après avoir utilisé le modèle RFM, il est obtenu un résultat de 29 785 clients. La table 5 présente un fragment de la base finale de données.
Dans cette approche, les données de la variable R (Récence) sont des entiers, représentant l'intervalle (en jours) entre le dernier achat du client et la date de fin de la période quand cette étude est effectuée. Cela signifie donc que le chiffre le plus petit correspond à la valeur client la plus grande.

Numéro client	Récence (R)	Fréquence (F)	Monétaire (M)
1	18	3	1 900 317
2	55	6	2 897 889
3	98	15	2 849 974
...	...	...	...
29 785	3	2	335 100

Table 5. Le résultat obtenu après avoir appliqué les variables RFM [22]

Concernant le système de notation, on peut constater que sur cette approche il ne s'appliquent pas des critères métiers (comme nous avons remarqué dans le système de [23]). Leur solution comprend une normalisation des données en utilisant la méthode Min-Max.
La normalisation Min-Max est une technique simple qui ajuste spécifiquement les données dans un intervalle prédéfini. [24]

$$A' = \left(\frac{A - valeur\ min\ A}{valeur\ max\ A - valeur\ min\ A}\right) \cdot (D - C) + C \qquad \text{[Equation 1] [24]}$$

Où,

- A - la valeur de donnée
- Valeur min/max A - la valeur min/max parmi les données
- [C, D] - l'intervalle des données

Voici la méthode de normalisation Min-Max utilisée dans l'exemple de l'article [22].

$$NR = \frac{R_{max} - R}{R_{max} - R_{min}} \qquad NF = \frac{F - F_{min}}{F_{max} - F_{min}} \qquad NM = \frac{M - M_{min}}{M_{max} - M_{min}}$$

Le résultat est présenté sur la table 6.

Cluster Number	NR	NF	NM
1	0.9143	0.0364	0.0943
2	0.7381	0.0909	0.1441
3	0.5333	0.2545	0.1417
...	...	...	...
29875	0.9857	0.0182	0.0161

Table 6. Le résultat après avoir appliqué la normalisation Min-Max[22]

Finalement, après avoir appliqué la méthode de la normalisation Min-Max, toutes les données sont dans la même intervalle, entre 0 - 1.

6.2. L'algorithme K-means

K-means est l'un des algorithmes de clustering [7]. Il prend comme paramètre le nombre de clusters et partitionne les données dans le nombre défini de clusters de sorte que la similarité entre ces derniers soit élevée [21]. K-Means est une approche itérative qui calcule la valeur des centroïdes avant chaque itération. Il nécessite des nombres précis de clusters **k**, car le centre du cluster initial peut changer, de sorte que cet événement peut entraîner un regroupement instable des données [25]. Les points de données sont déplacés entre différents clusters en fonction des centroïdes calculés à chaque itération [21]. Le processus est répété jusqu'à ce que la somme ne puisse plus être diminuée.
Les avantages de K-means incluent sa vitesse et son évolutivité : c'est l'un des modèles de clustering les plus rapides et il peut gérer efficacement des ensembles de données longs et larges avec de nombreux enregistrements et de nombreux champs de clustering d'entrée.[7]

L'algorithme K-Means est présenté dans l'algorithme 1 .

1. Initialement, en fonction de la valeur de **k**, **k** points aléatoires sont choisis comme centroïdes initiaux.
2. Les distances de chaque point de données aux centroïdes choisis précédemment sont évaluées à l'aide de la distance euclidienne.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

3. Les valeurs de distance sont comparées et le point de données est attribué au centroïde qui a la valeur de distance euclidienne la plus courte.
4. Les étapes précédentes sont répétées. Le processus est arrêté si les clusters obtenus sont les mêmes que ceux de l'étape précédente.

[Algorithme 1] [21]

Les étapes de l'algorithme sont présentées par Lebart et al., 1995 [26].

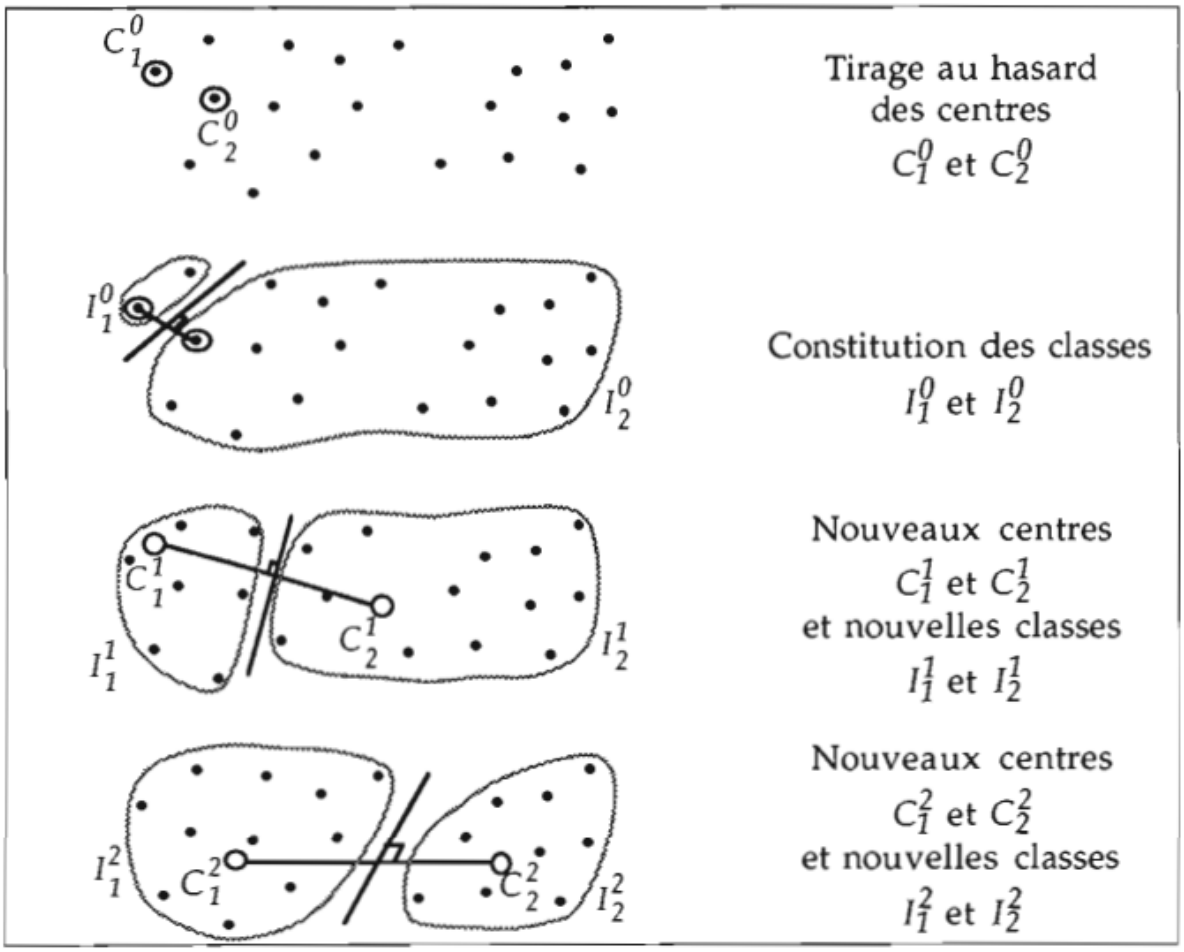

Figure 7. Les étapes de l'algorithme K-means [26]

6.2.1. Comment trouve-t-on le nombre optimal de clusters K ?

La sélection de la valeur de **K** comme nombre de clusters a un grand impact sur les résultats de clustering [28].

Le projet de Data Mining doit commencer par une compréhension de l'objectif métier et une évaluation de la situation actuelle ainsi que des problèmes [10].
Pour cette raison, le nombre de clusters K peut être défini par le niveau opérationnel de l'entreprise (marketeurs, analystes de business) en considérant leur stratégie marketing.[4][11]

La stratégie marketing peut inclure le ciblage d'une partie spécifique de clients [4] et cela changerait donc le nombre de clusters. Également, les responsables des systèmes de fidélités sont généralement des agences spécialisées dans le conseil en fidélisation, la créativité, la communication, l'analyse de données, les logiciels de fidélisation et les opérations back-end.
Ils forment généralement des segments de clientèle basés sur les pratiques de l'industrie, les connaissances accumulées et les méthodes statistiques d'analyse des données. [11]

Il existe une autre méthode pour déterminer la meilleure valeur de **K**.
La méthode Elbow représente une courbe de relation entre la somme des carrés de l'erreur (SCE) et **K.** Cette courbe prend la forme d'un coude, et la valeur de **K** correspondant à ce coude est le véritable numéro de cluster des données. [28]
L'indicateur de base de la méthode Elbow est la SCE (la somme des carrés de l'erreur), comme indiqué dans l'équation suivante :

$$SCE = \sum_{j=1}^{k} \sum_{i}^{n} | x_i^{(j)} - c_j |^2 \qquad \text{[Equation 2] [29]}$$

Où :

- k - le nombre de clusters
- n - le nombre d'objets dans un cluster
- x - l'élément dans le cluster
- c_j - le centroid du cluster

Sur la figure 8, il est présenté un exemple de la courbe Elbow [28]. A en juger par la méthode du coude (voir figure 8), la diminution de SCE n'est pas significative lorsque K est supérieur à 4. Par conséquent, dans cet exemple, choisir K = 4 donnerait un résultat favorable.

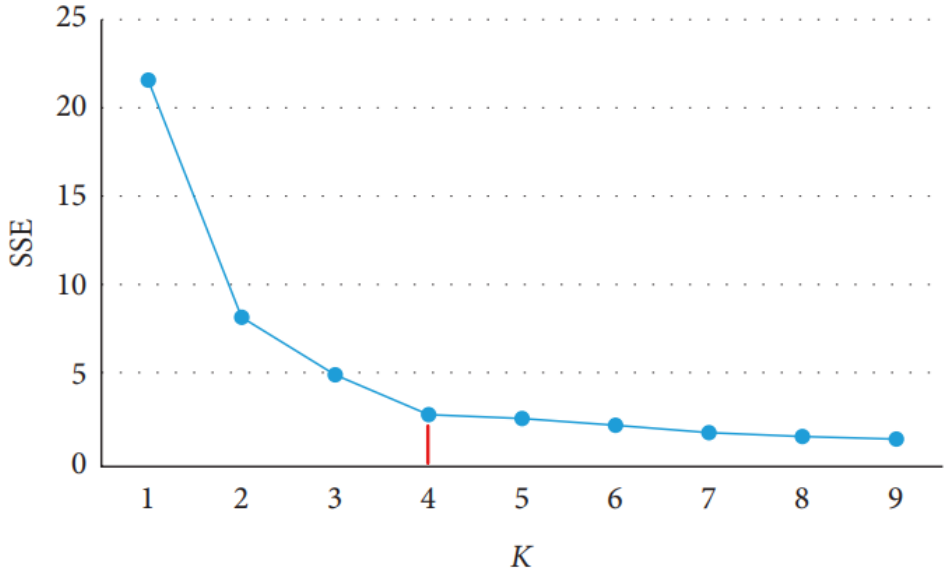

Figure 8. Résultat du nombre optimal de clusters d'utilisateurs avec la méthode Elbow [28]

6.2.2. L'application de l'algorithme K-means

Dans l'exemple [28], les auteurs ont effectué un experiment sur un ensemble de données de 10 248 transactions, créées sur une plateforme d'achat du 1er Novembre 2017 au 15 avril 2019. Cette base de données implique 1 013 clients.

Après avoir appliqué le modèle RFM ainsi que la méthode Elbow pour trouver le nombre optimal de clusters (voir figure 8) ils ont implémenté l'algorithme K-means sur Python.

Voici une présentation visuelle du résultat. (voir figure 9)

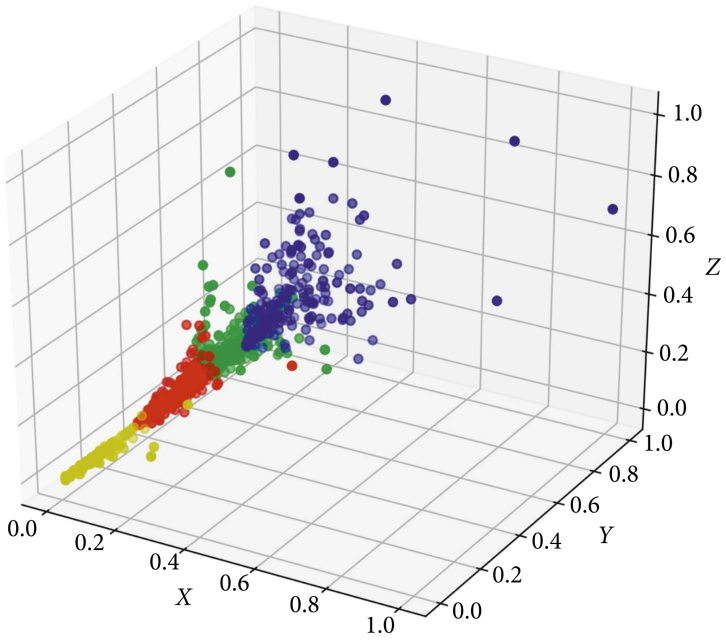

Figure 9. Scatterplot basé sur l'algorithme K-means [28]

Dans le graphique, l'axe X représente le montant total des achats (la variable monétaire dans le modèle RFM), l'axe Y représente la date d'achat la plus récente (R) et l'axe Z représente la fréquence d'achat (F). On peut voir que les données globales de l'utilisateur sont proches de 0 sur l'axe X.

Les indicateurs des différents groupes de clients et ceux de l'ensemble des clients sont également pertinents pour l'analyse suivante. Ils sont présentés sur la table 7.

Le groupe de client	Le temps moyen depuis le dernier achat	Moyenne du montant total des achats	La fréquence d'achat moyenne	Le nombre de clients
Groupe 1	457.27	79.86	2.88	98
Groupe 2	46.93	452.15	25.47	184
Groupe 3	262.1	65.29	4.55	415
Groupe 4	157.54	156.25	10.71	316
Total	209.3	165.34	10.11	1013

Table 7. Les indicateurs des clients [28]

Le graphique (voir figure 9) montre qu'un petit nombre de clients dépasse de loin le montant moyen des achats. Leur ID utilisateur et les enregistrements d'achat correspondants sont extraits comme indiqué dans la table 8.

Employee_id	Price_total	Last_time	Frequency
YH170317000002	3167.5	2019/3/26 14 :33	86
YH171201000030	2394.5	2019/1/21 14 :01	108
YH180212000003	2251.5	2019/1/21 11 :22	44
YH171116000009	1513	2019/1/26 22 :54	37

Table 8. Les données extraites des valeurs aberrantes [28]

Nous remarquons que ces clients ont un montant total d'achat et une fréquence d'achat plus élevée. Ils ont également un temps plus court depuis leur dernier achat. Pour cette raison, le résultat combiné est une meilleure performance sur les trois axes et ils sont représentés comme des valeurs aberrantes sur le scatter plot.

Les quatre clients de la table 8 peuvent être attribués au groupe 2, qui se caractérise par un montant total d'achat important et une fréquence d'achat élevée. Les données de ces utilisateurs sont donc alignées avec les caractéristiques globales de ce groupe, prouvant ainsi la rationalité du regroupement.

Les clients du groupe 1 ont plus de temps depuis le dernier achat. D'ailleurs, leur montant total d'achat et la fréquence sont assez faibles. Ces clients peuvent être considérés comme des clients qui présentent des risques de perte et qui nécessitent une observation plus approfondie.

La fréquence d'achat et le montant total des achats des clients du groupe 2 sont supérieurs aux moyennes globales, et leur dernier achat est également plus récent, ce qui indique qu'ils sont des clients d'une grande valeur. Selon [28], la plateforme devrait faire plus d'efforts pour maintenir et améliorer la relation avec eux.

Le montant total des achats et la fréquence d'achat des clients du groupe 3 sont faibles et ces clients ont effectué leur dernier achat plus tôt que la moyenne. Cela implique que malgré leur comportement d'achat récent sur la plateforme, ils n'y ont pas pris d'habitude de consommation et ne sont pas capables de générer de gros bénéfices pour la plateforme.
On remarque que le nombre des clients du groupe 3 est très élevé. Ils peuvent donc être considérés comme des clients types. La plateforme doit cultiver leurs habitudes et essayer de les convertir en clients actifs qui peuvent apporter plus de bénéfices.

Les clients du groupe 4 ont effectué leur dernier achat à une date relativement récente. Leurs indicateurs de montant total d'achat et de fréquence d'achat sont proches des moyennes globales. On dirait que ces clients sont plus actifs et ont formé certaines habitudes de consommation sur la plateforme. Notamment, il reste encore beaucoup à faire pour améliorer leur montant total d'achat et leur fréquence d'achat, ce qui signifie qu'ils doivent être traités comme des clients à fort potentiel. Les objectifs marketing de la plate-forme pour eux devraient se concentrer sur leur passage du groupe 4 au groupe 2 avec une fréquence et un montant d'achat plus élevé.

7. Une amélioration de l'algorithme K-means

7.1. Les inconvénients de l'algorithme K-means

La méthode K-Means est la méthode de clustering la plus simple et la plus connue [23][25].
Cependant, K-Means présente certains inconvénients, comme suit :

- K-Means Clustering est une méthode d'optimisation localisée qui est sensible à la sélection de la position de départ à partir du milieu du cluster [25].
- Le nombre d'itérations de l'algorithme est affecté par le centroïde initial du cluster au hasard [25].
- Le centroïde du cluster peut se regrouper plus près l'un de l'autre, ce qui rend les clusters moins significatifs [21].
- Le nombre de clusters doit être prédéterminé. Pourtant, dans plusieurs cas, il est difficile de prédéterminer le nombre de clusters [30].
- Il est inadéquat pour regrouper des variables qualitatives [30].

7.2. Une version améliorée de l'algorithme K-means

Dans [30], les auteurs présentent une approche améliorée de l'algorithme K-means, appelée "K-means à trois niveaux" (tri-level K-means). Ce dernier est un algorithme qui fournit un ensemble plus correct de centroïdes initiaux et réduit les problèmes de données anormales ou aberrantes. Il peut donc fournir un meilleur résultat de clustering que l'algorithme K-means traditionnel.
L'algorithme K-means traditionnel attribue une donnée P au cluster dont la distance entre P et le centre du cluster est minimale [26][30]. De cette manière, les données aberrantes ou bruitées peuvent être attribuées à des clusters contenant moins de données et en même temps, les données normales ne sont attribuées qu'à quelques clusters contenant chacun beaucoup de données.
Par exemple, sur la figure 10, il y a deux données anormales indiquées par des points rouges et bleus qui sont classées en deux groupes. Par contre, les données normales indiquées par des

points noirs sont classées en un seul groupe. Un traitement ultérieur n'est généralement pas utile dans cette situation [30].

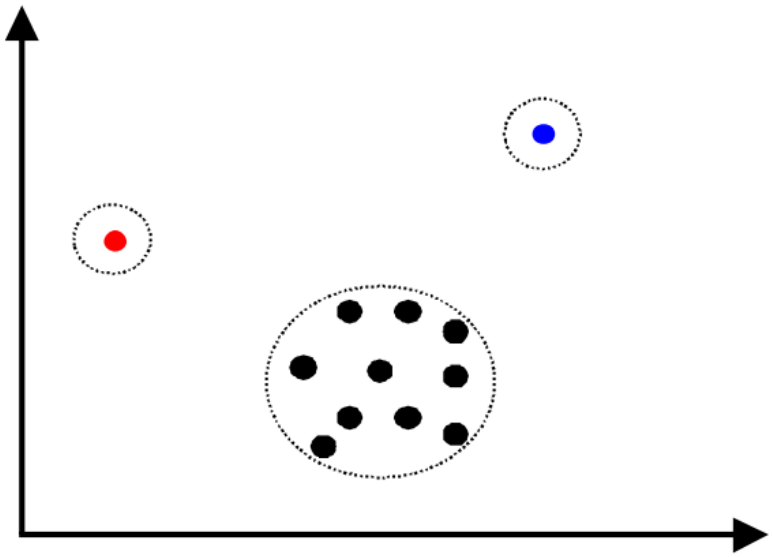

Figure 10. Exemple de l'algorithme K-means traditionnel [30]

Dans [30], l'algorithme K-means a trois niveaux est divisé en quatre étapes :

1. **La normalisation des données**

Les différentes unités de mesure des variables affectent les résultats de clustering, pour cette raison dans [30], les auteures insistent que la normalisation des données est indispensable avant d'appliquer l'algorithme. Une méthode très utilisée afin de résoudre cet inconvénient est la technique Min-Max qu'on a aussi traité sur la partie 6.1.1, exemple de l'article [22].

La normalisation Min-Max est une technique simple qui ajuste spécifiquement les données dans un intervalle prédéfini [24].

Considérons l'ensemble de données $S' = \{X'_1, X'_2, X'_3, \dots X'_N\}$, composé de N données et de $\boldsymbol{d}$ nombre de dimensions où $(X'_{i1}, X'_{i2}, X'_{i3}, \dots\dots, X'_{id})$ sont les variables de X'_i. En utilisant l'équation 2, chaque valeur de variable X'_{ij} est normalisee en x_{ij}. Soit X_i constitue de $(x_{i1}, x_{i2} \dots., x_{id})$ la donnée normalisée de X'_i. On obtient donc $S = \{X_1, X_2, X_3, \dots X_N\}$.

$$x_{ij} = \frac{x'_{ij} - MIN(x'_{kj})}{MAX(x'_{kj}) - MIN(x'_{kj})} \qquad \forall\, k \in [1,\ N]$$ [Equation 2] [30]

2. Clustering de premier niveau

Dans cette première étape, l'algorithme K-means à trois niveaux classe les données qui se trouvent dans S en K' grands clusters où $K' < K$, en appliquant l'algorithme K-means traditionnel. En analysant la quantité et la variation des données dans chaque grand groupement, on peut obtenir le nombre de sous-clusters qui peuvent être dérivés. Pour cette raison, cet algorithme amélioré, présenté dans l'article [30], divise d'abord les données en grands segments.

3. Clustering de deuxième niveau

L'objectif de cette étape est de diviser un grand cluster en plusieurs sous-clusters. Les facteurs analysés sont la quantité du cluster et le niveau d'écart entre ses données. Dans [30], la mesure utilisée afin de calculer l'écart des données est la déviation standard. Par conséquent, le grand cluster avec une plus grande quantité et une déviation standard élevée, sera divisé en plus petits clusters.

Soit (x_{ci1}, x_{ci2}, x_{ci3}, ..., x_{cid}) la ***i***-ème donnée dans le ***c***-ième grand clusters obtenu à l'étape de clustering de premier niveau. Le but de cette phase est de diviser les K' grand clusters en K clusters où $K = K_1 + K_2 + \ldots + K_{K'}$.

Soit n_c le nombre de données dans le ***c***-ième grand cluster. La déviation standard de ce cluster sera calculée par l'équation 3.

$$Std_c = \frac{\sum_{j=1}^{d} \sqrt{\frac{\sum_{i=1}^{n_c} (x_{cij} - u_{cj})^2}{n_c}}}{d}, \quad \text{où} \quad u_{cj} = \frac{\sum_{i=1}^{n_c} x_{cij}}{n_c}.$$

[Equation 3] [30]

Ensuite, l'algorithme K-means à trois niveaux classe les données de chaque grand cluster ***c*** en clusters K_c par un algorithme k-means traditionnel, où K_c est défini comme suit :

$$K_c = \frac{n_c \times Std_c}{\sum_{i=1}^{K'} (n_i \times Std_i)} \times K$$ [Equation 4] [30]

4. Clustering de niveau final

Dans l'étape précédente on a obtenu comme résultat K clusters $(C_1', C_2', C_3', \ldots, C_K')$.

Soit $(x_{ci1}, x_{ci2}, \ldots x_{cid})$ les variables de ***i***-ème donnée du C_C' cluster et n_c' le nombre total de données dans C_C'.

Pour chaque cluster on obtient son centre c_{cj} et $(c_{c1}, c_{c2}, \ldots, c_{cd})$ (voir équation 5).

$$c_{cj} = \frac{1}{n_c'} \sum_{i=1}^{n_c'} x_{cij}$$ [Equation 5] [30]

Dans cette étape, l'algorithme attribue les centres de clusters $C_1', C_2', \ldots C_K'$ tant que centres initiaux et applique le K-means traditionnel pour classer les données dans S en K clusters.

Sur la figure 11 et 12 est illustré l'algorithme K-means traditionnel et l'algorithme K-means à trois niveaux respectivement. Les points noirs sont les données et les points rouges représentent les centres des clusters.

La figure 11 illustre les étapes de l'algorithme K-means dans le pire des cas où les centroïdes initiaux sont près l'un de l'autre. On remarque que le résultat obtenu n'est pas optimal, une situation similaire de celui présenté sur la figure 10.

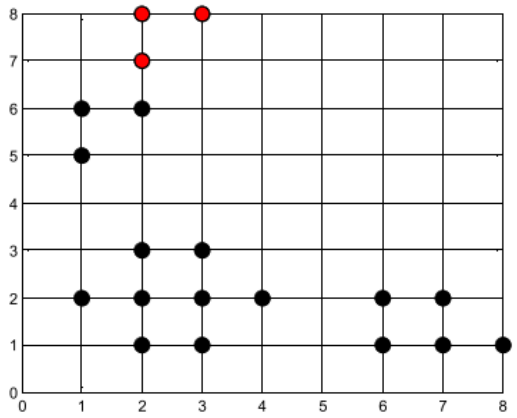

Figure 11. a. Les données et les centres initiaux [30]

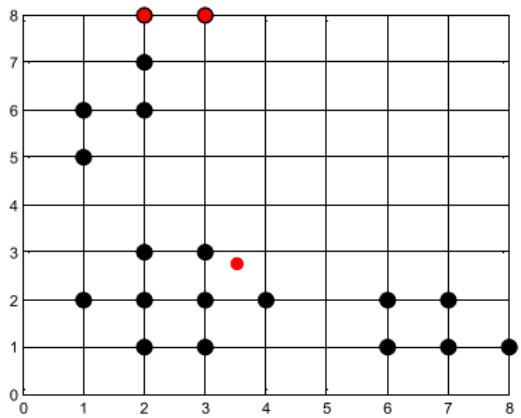

Figure 11. b. Le résultat après la première itération [30]

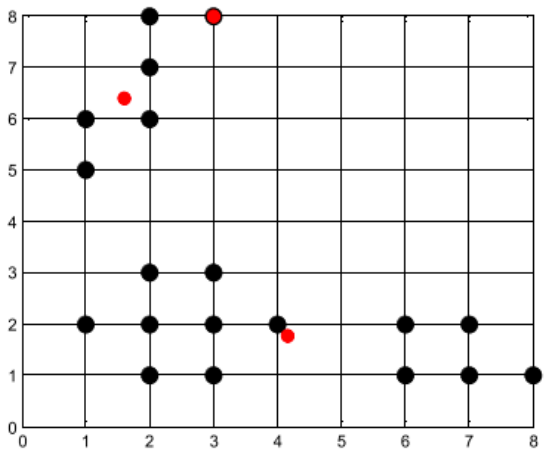

Figure 11. c. Le résultat après la deuxième itération [30]

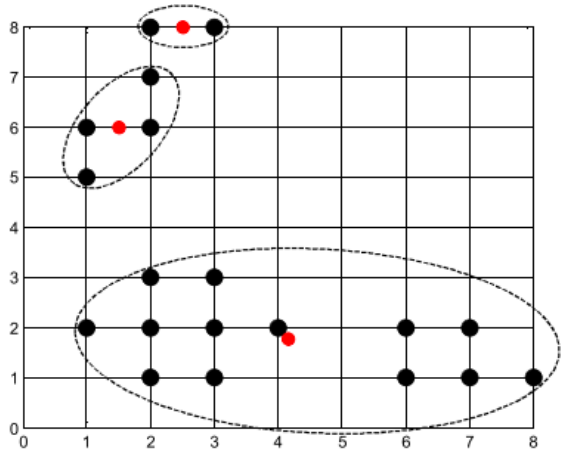

Figure 11. d. Le résultat final après la troisième itération [30]

La figure 12 illustre les étapes de l'algorithme K-means à trois niveaux dans le pire des cas où les centroïdes initiaux sont près l'un de l'autre.

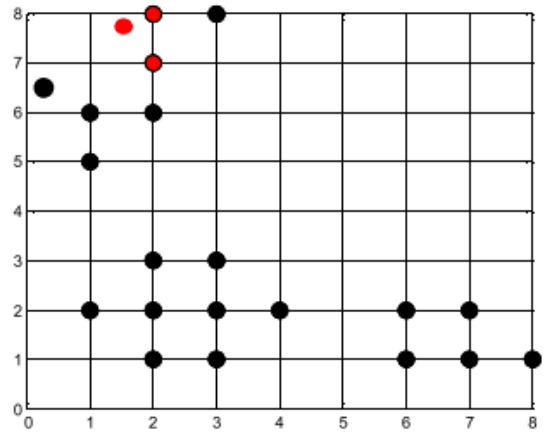

Figure 12. a. Les données et les centres initiaux [30]

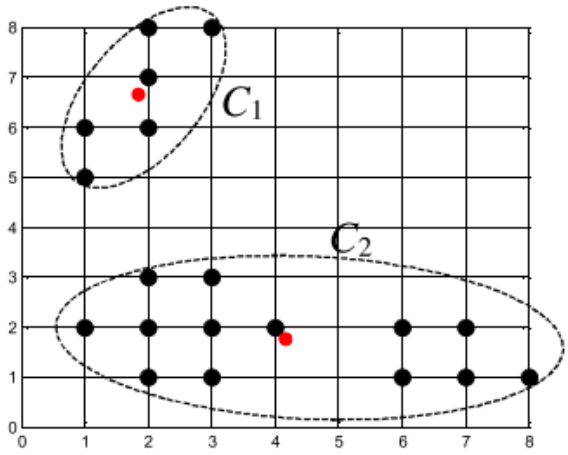

Figure 12. b. Le résultat après le clustering de premier niveau [30]

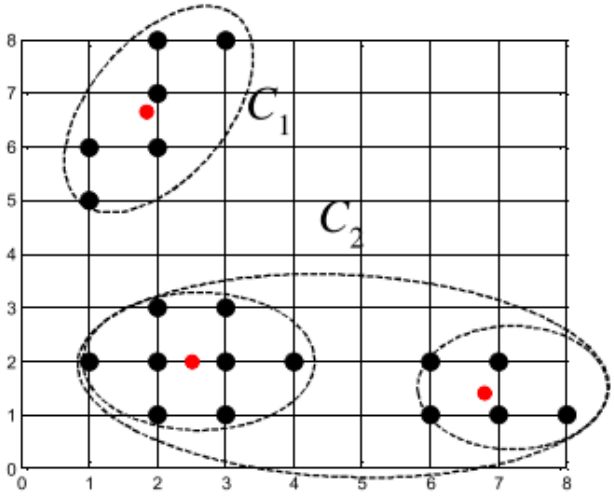

Figure 12. c. Le résultat après le clustering de deuxième niveau. [30]

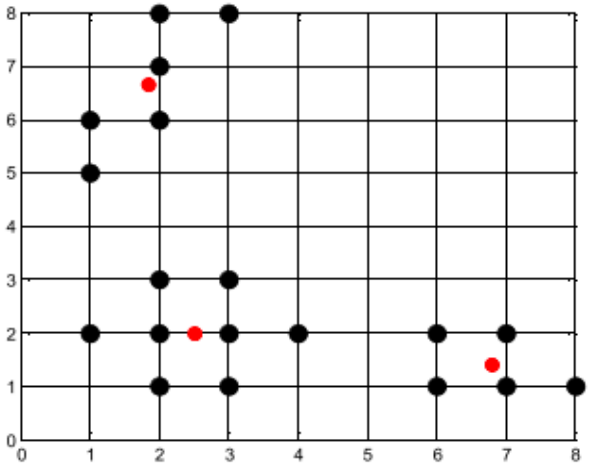

Figure 12. d. Les données et les centres utilisés dans l'étape de clustering de niveau final. [30]

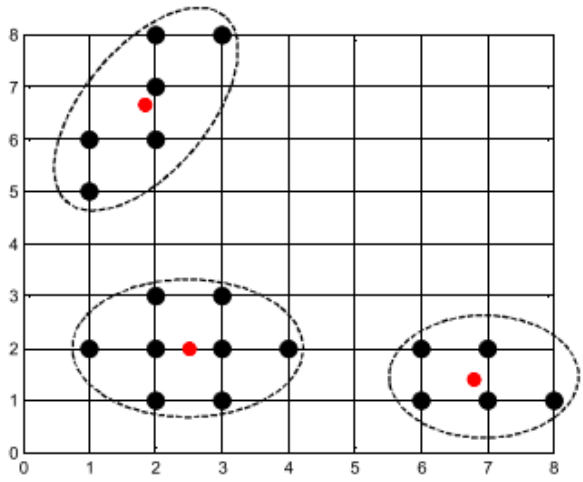

Figure 12.e. Le résultat obtenu après le clustering de niveau final [30]

8. Conclusion

L'objectif de cet état de l'art est de présenter l'importance de la segmentation client ainsi que l'utilisation de Data Mining afin de grouper la clientèle d'une façon pertinente pour des missions métiers.

Nous avons montré que la segmentation de la clientèle, une partie importante de CRM, est appliquée dans plusieurs secteurs du marché. L'objectif principal de ce processus est de créer certains groupes clients en se basant sur leurs derniers achats. De cette manière, les clients d'un comportement d'achat similaire appartiennent au même segment. Ce résultat peut être utilisé par le niveau managérial de l'entreprise afin de lancer des campagnes marketing différentes pour chaque groupe client.

Ensuite, nous avons parcouru les techniques descriptives et prédictives de Data Mining. Ce dernier est un domaine très large et il est toujours en cours de développement.
Nous avons concentré notre recherche spécifiquement sur l'application de Data Mining dans le but de créer des segments clients.
Il est remarqué l'importance d'un processus bien défini afin de traiter les données, un processus qui peut varier des types de données, du secteur et d'objectifs métiers.

A travers cette recherche nous avons observé un processus appliqué majoritairement qui s'agit d'un modèle RFM suivi par le clustering. L'algorithme K-means est l'algorithme le plus connu du clustering, car il est facile à implémenter et il peut traiter un grand nombre de données rapidement. Toutefois, plusieurs articles de recherche présentent certains inconvénients de K-means concernant le nombre de clusters K, qui doit être prédéfini ainsi que les centroïdes initiaux qui sont choisis d'une manière aléatoire.
Pour cette raison, nous avons trouvé pertinent d'élargir notre recherche afin de trouver des solutions qui surmontent ces problèmes. Le nombre optimal de clusters K peut être obtenu par la méthode Elbow (voir figure 7), mais aussi il peut être défini par les marketeurs en considérant leur stratégie marketing. Au cours de ce travail de recherche nous avons remarqué que le nombre d'études concernant l'amélioration de l'algorithme K-means a connu une grande augmentation ces

dernières années. Nous avons étudié une version améliorée de cet algorithme, "L'algorithme K-means à trois niveaux". Ce dernier prend un nombre de clusters inférieur à K et en analysant la déviation standard des données, il divise chaque grand cluster en sous-clusters.

En conclusion, le Data Mining est représenté en tant qu'un outil utilisé afin de réaliser la segmentation de la clientèle. En considérant le fait qu'on traite des données comportementales, il n'est pas toujours assez explicite de bien définir les variables comportementaux dans un marché très dynamique. D'ailleurs les stratégies marketing se développent rapidement et elles sont de plus en plus orientées vers des campagnes personnalisées. Pour cette raison, des études supplémentaires sont nécessaires afin d'analyser les meilleurs outils pour une stratégie marketing alignée aux exigences du marché.

Références

[1] Parvatiyar, Atul, and Jagdish N. Sheth. "Customer relationship management : Emerging practice, process, and discipline." Journal of Economic & Social Research 3, no. 2 (2001).

[2] Rababah, Khalid, Haslina Mohd, and Huda Ibrahim. "Customer relationship management (CRM) processes from theory to practice : The pre-implementation plan of the CRM system." International Journal of e-Education, e-Business, e-Management and e-Learning 1, no. 1 (2011) : 22-27.

[3] A. H. Kracklauer, D. Q. Mills, and D. Seifert. Customer management as the origin of collaborative customer relationship management. pages 3–6, 2004.

[4] Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018, October). Customer segmentation and strategy development based on user behavior analysis, RFM model and data mining techniques : a case study. In 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE) (pp. 119-126). IEEE.

[5] Smith, Wendell R. "Product differentiation and market segmentation as alternative marketing strategies." Journal of marketing 21, no. 1 (1956) : 3-8.

[6] Danaher, Peter J. "Customer heterogeneity in service management." Journal of Service Research 1, no. 2 (1998) : 129-139.

[7] Tsiptsis, Konstantinos K., and Antonios Chorianopoulos. "Data mining techniques in CRM : inside customer segmentation." (2011).

[8] Khalili-Damghani, Kaveh, Farshid Abdi, and Shaghayegh Abolmakarem. "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem : Real case of customer-centric industries." Applied Soft Computing 73 (2018) : 816-828.

[9] Wang, Shu-Ching, Yao-Te Tsai, and Yi-Syuan Ciou. "A hybrid big data analytical approach for analyzing customer patterns through an integrated supply chain network." Journal of Industrial Information Integration 20 (2020) : 100177.

[10] Ziafat, Hasan, and Majid Shakeri. "Using data mining techniques in customer segmentation." Journal of Engineering Research and Applications 4, no. 9 (2014) : 70-79.

[11] Mihova, Vesela, and Velisar Pavlov. "A customer segmentation approach in commercial banks." In AIP conference proceedings, vol. 2025, no. 1, p. 030003. AIP Publishing LLC, 2018.

[12] Yi, Youjae, and Hoseong Jeon. "Effects of loyalty programs on value perception, program loyalty, and brand loyalty." Journal of the academy of marketing science 31, no. 3 (2003) : 229-240.

[13] Dogan, Onur, Ejder Ayçin, and Zeki Bulut. "Customer segmentation by using RFM model and clustering methods : a case study in retail industry." International Journal of Contemporary Economics and Administrative Sciences 8 (2018).

[14] Ballestar, María Teresa, Pilar Grau-Carles, and Jorge Sainz. "Customer segmentation in e-commerce : Applications to the cashback business model." Journal of Business Research 88 (2018) : 407-414.

[15] Christino, Juliana Maria Magalhães, Thaís Santos Silva, Erico Aurélio Abreu Cardozo, Alexandre de Pádua Carrieri, and Patricia de Paiva Nunes. "Understanding affiliation to cashback programs : An emerging technique in an emerging country." Journal of Retailing and Consumer Services 47 (2019) : 78-86.

[16] Maksood, Fathimath Zuha, and Geetha Achuthan. "Analysis of data mining techniques and its applications." International Journal of Computer Applications 140, no. 3 (2016) : 6-14.

[17] Chen, Feng, Pan Deng, Jiafu Wan, Daqiang Zhang, Athanasios V. Vasilakos, and Xiaohui Rong. "Data mining for the internet of things : literature review and challenges." International Journal of Distributed Sensor Networks 11, no. 8 (2015) : 431047.

[18] Cheng, Ying, Ken Chen, Hemeng Sun, Yongping Zhang, and Fei Tao. "Data and knowledge mining with big data towards smart production." Journal of Industrial Information Integration 9 (2018) : 1-13.

[19] Wright, Aileen P., Adam T. Wright, Allison B. McCoy, and Dean F. Sittig. "The use of sequential pattern mining to predict next prescribed medications." Journal of biomedical informatics 53 (2015) : 73-80.

[20] Ramamohan, Y., K. Vasantharao, C. Kalyana Chakravarti, and A. S. K. Ratnam. "A study of data mining tools in the knowledge discovery process." International Journal of Soft Computing and Engineering (IJSCE) 2, no. 3 (2012) : 191-194.

[21] Christy, A. Joy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa. "RFM ranking–An effective approach to customer segmentation." Journal of King Saud University-Computer and Information Sciences 33, no. 10 (2021) : 1251-1257.

[22] Ahang, Farahnaz, Abdolmajid Imani, Meysam Abbasi, Hassan Ghaffari, and Mohamad Mehdi. "Customer segmentation to identify key customers based on RFM model by using data mining techniques." International journal of research in industrial engineering 11, no. 1 (2022) : 62-76.

[23] Maryani, Ina, Dwiza Riana, Rachmawati Darma Astuti, Ahmad Ishaq, and Eva Argarini Pratama. "Customer segmentation based on RFM model and clustering techniques with K-means algorithm." In 2018 Third International Conference on Informatics and Computing (ICIC), pp. 1-6. IEEE, 2018.

[24] Patro, S., and Kishore Kumar Sahu. "Normalization : A preprocessing stage." arXiv preprint arXiv :1503.06462 (2015).

[25] "Syakur, M. A., B. K. Khotimah, E. M. S. Rochman, and Budi Dwi Satoto. ""Integration k-means clustering method and elbow method for identification of the best customer profile cluster."" In IOP conference series : materials science and engineering, vol. 336, no. 1, p. 012017. IOP Publishing, 2018."

[26] Lebart, Ludovic, Alain Morineau, and Marie Piron. Statistique exploratoire multidimensionnelle. Vol. 3. Paris : Dunod, 1995.

[27] Thakare, Y. S., and S. B. Bagal. "Performance evaluation of K-means clustering algorithm with various distance metrics." International Journal of Computer Applications 110, no. 11 (2015) : 12-16.

[28] Wu, Jun, Li Shi, Wen-Pin Lin, Sang-Bing Tsai, Yuanyuan Li, Liping Yang, and Guangshu Xu. "An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm." Mathematical Problems in Engineering 2020 (2020).

[29] Gustriansyah, Rendra & Suhandi, Nazori & Antony, Fery. (2020). Clustering optimization in RFM analysis Based on k-Means. Indonesian Journal of Electrical Engineering and Computer Science. 18. 470. 10.11591/ijeecs.v18.i1.pp470-477.

[30] Yu, Shyr-Shen, Shao-Wei Chu, Chuin-Mu Wang, Yung-Kuan Chan, and Ting-Cheng Chang. "Two improved k-means algorithms." *Applied Soft Computing* 68 (2018) : 747-755.