Titre: Title:	Développement d'un modèle de prédiction de retours de vêtements en fonction des caractéristiques des clients et des produits
Auteur: Author:	Whitney Kate Chucya Lozano
Date:	2020
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Chucya Lozano, W. K. (2020). Développement d'un modèle de prédiction de retours de vêtements en fonction des caractéristiques des clients et des produits [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/5519/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/5519/
Directeurs de recherche: Advisors:	Bruno Agard
Programme: Program:	Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL

Affiliée à l'Université de Montréal

Développement d'un modèle de prédiction de retours de vêtements en fonction des caractéristiques des clients et des produits

WHITNEY KATE CHUCYA LOZANO

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Novembre 2020

POLYTECHNIQUE MONTRÉAL

Affiliée à l'Université de Montréal

Ce mémoire intitulé :

Développement d'un modèle de prédiction de retours de vêtements en fonction des caractéristiques des clients et des produits

présenté par **Whitney Kate CHUCYA LOZANO**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Christophe DANJOU, président

Bruno AGARD, directeur de recherche

Benoît ROBERT, membre

DÉDICACE

Pleine de joie, je dédie ce travail à chacun de mes proches, qui ont été mon soutien pour réussir. C'est une grande satisfaction pour moi de pouvoir le leur dédier, ce que j'ai accompli avec beaucoup d'efforts et de travail.

À mes parents bien-aimés Edgar et Magdalena, car ils sont la motivation de ma vie et ma fierté d'être ce que je suis.

À mon frère Kendall et à ma belle-sœur Gleny, pour m'avoir donné une raison de plus pour atteindre mes objectifs.

À M. Bruno Agard, pour m'avoir fait confiance et m'avoir guidé pendant tout ce temps

À mon cher Corentin, pour son soutien, sa compréhension et son amour dans les moments difficiles.

À tous ceux que j'aime, merci

Whitney

REMERCIEMENTS

La réussite de ce travail de recherche a été possible non seulement grâce à mes efforts, mais aussi grâce au soutien de plusieurs personnes, qui m'ont donné leurs conseils et ont gardé un œil sur mes progrès pendant ma formation académique.

Dans ces lignes, je tiens à remercier toutes les personnes qui ont rendu cette recherche possible et qui ont été avec moi. Tout d'abord, je tiens à exprimer ma gratitude à mon directeur de recherche, M. Bruno Agard, qui a guidé mon travail avec ses connaissances. Merci de m'avoir donné l'opportunité de faire partie de votre équipe, ainsi que d'avoir eu la plus grande patience du monde pour me guider et me soutenir dans le développement de cette recherche.

Je remercie mes collègues du Laboratoire en intelligence de données de m'avoir accueilli et soutenu lorsque j'en ai eu besoin. Merci de m'avoir faite sentir chez moi depuis mon arrivée et d'avoir partagé vos connaissances avec moi.

Je tiens également à remercier notre partenaire Logistik Unicorp. J'apprécie la confiance qui m'a été accordée pour mener à bien cette étude. Merci de m'avoir fourni les ressources et les outils nécessaires.

Je remercie mes parents pour leur temps, leur amour, leurs enseignements et les sacrifices qu'ils ont faits pour moi. À mon frère, qui a toujours été comme un deuxième père et mon modèle. À ma cousine Ana Lucia qui est ma meilleure amie et qui ne m'a laissé jamais seule malgré la distance.

Et bien sûr à l'École Polytechnique de Montréal, à mes enseignants et aux membres du personnel administratif, pour m'avoir permis de conclure une étape de ma vie, merci pour vos conseils et m'avoir guidé tout au long de cette recherche.

RÉSUMÉ

Dans l'industrie manufacturière, le processus de gestion des retours de produits est crucial. Quelque soit le volume des ventes d'une entreprise, les retours sont toujours un problème pour les commerçants puisqu'un retour coûte non seulement de l'argent, mais aussi du temps. De plus, ces dernières années, la croissance accélérée du commerce électronique a rendu le problème encore plus difficile, en particulier dans l'industrie du vêtement. Notre projet est basé sur une étude de cas au sein d'une entreprise qui se charge de la sous-traitance des programmes d'uniformes. L'entreprise fournit à ses clients une plateforme d'achat en ligne pour acheter des pièces d'uniformes.

Dans ce mémoire, on utilise des méthodes d'apprentissage supervisées et non supervisées pour analyser l'historique d'achat et de retours de produits au cours d'une période d'étude spécifique afin de trouver les clients qui sont les plus susceptibles d'effectuer des retours, ainsi que des produits problématiques qui sont souvent retournés. Tout d'abord, il a fallu transformer les données disponibles en séries chronologiques pour pouvoir observer l'évolution des mesures morphologiques des clients dans le temps. Quatre mesures principales ont été analysées, en tenant compte des vêtements avec le pourcentage le plus élevé de commandes et de retours. Ensuite, nous avons procédé à chercher de possibles comportements similaires entre les clients en utilisant une méthode d'apprentissage non supervisée, la segmentation. La segmentation consiste à regrouper un ensemble d'objets en sous-ensembles d'objets appelés clusters, où chaque cluster est constitué d'une collection d'objets qui sont similaires les uns aux autres, mais qui sont différents des objets des autres clusters. On a utilisé la métrique appelée Dynamic Time Warping car elle est la plus adaptée pour comparer des séries temporelles. De cette façon, nous avons été capables d'obtenir des groupes de clients ayant des comportements similaires. Avec les résultats obtenus, on a procédé à l'élaboration d'un modèle qui nous permet de reconnaître les caractéristiques des clients qui effectuent des retours. De plus, un modèle capable de montrer les spécifications des produits qui sont toujours retournés a été développé.

Les algorithmes d'apprentissage basés sur des arbres de décision sont considérés comme l'une des méthodes d'apprentissage supervisé les meilleures et les plus largement utilisées. Les méthodes basées sur les arbres de décision renforcent les modèles prédictifs avec une précision, une stabilité et une facilité d'interprétation élevées. Dans notre cas, on utilise les arbres de classification, qui

possèdent un processus en deux étapes, une étape d'apprentissage et une étape de prédiction. Dans l'étape d'apprentissage, le modèle est développé sur une base des données d'entraînement. Tandis que dans l'étape de prédiction, le modèle est utilisé pour prédire si un retour sera effectué ou non.

Différents tests ont été effectués sur la base de différentes méthodes et algorithmes, de plus on a utilisé des processus de « boosting » et « winnowing » pour améliorer les performances de nos arbres. Grâce aux résultats, on a été en mesure de trouver les caractéristiques et comportements de clients qui font des retours et on a constaté que même les clients avec des courbes d'évolution morphologique stables peuvent avoir des pourcentages de retours relativement élevés.

ABSTRACT

In the manufacturing industry, the process of managing product returns is crucial. Regardless of a company's sales volume, returns are always a problem for merchants as a return costs not only money, but also time. Moreover, in recent years, the accelerated growth of electronic commerce has made the problem even more difficult, especially in the clothing industry. This project is based on a case study in a company that outsources uniform programs. The company provides its customers with an online shopping platform to purchase their clothes.

In this paper, supervised and unsupervised learning methods are used to analyze the purchase history and product returns during a specific study period in order to find the customers who are most likely to make returns, as well as problematic products that are often returned. First of all, we had to transform the available data into time series in order to observe the evolution of the morphological measures of our clients over time. Four main metrics were analyzed, taking into account the garments with the highest percentage of orders and returns. Next, we proceeded to look for possible similar behaviors between clients using an unsupervised learning method, segmentation. Segmentation involves grouping a set of objects into subsets of objects called clusters, where each cluster is made up of a collection of objects that are similar to each other, but that are different from objects in other clusters. We used the metric called Dynamic Time Warping because it is the most suitable for comparing time series. Thus, we were able to get groups of customers with similar behaviors. With the results obtained, we proceeded to develop a model that allows us to recognize the characteristics of customers who make returns. In addition, a model capable of showing the specifications of products that are still returned has been developed.

Decision tree-based learning algorithms are considered to be one of the best and most widely used supervised learning methods. Decision tree-based methods enhance predictive models with high precision, stability, and ease of interpretation. In our case, we use classification trees, which have a two-step process, a learning step and a prediction step. In the training step, the model is developed on the basis of the training data. While in the prediction step, the model is used to predict whether or not a return will be made.

Different tests were carried out on the basis of different methods and algorithms. In addition, we used the processes of “boosting” and “winnowing” in order to improve the performance of our trees. Thanks to the results, we were able to find the characteristics and behaviors of clients who

return and it was found that even clients with stable morphological evolution curves may have relatively high return percentages.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

ARIMA : Autoregressive integrated moving average

DLM : Distributed lag model

DTW : Dynamic Time Warping

TADPole : Time-series Anytime Density Peaks

CHAPITRE 1 INTRODUCTION

L’extension de l’Internet et de sa technologie ces dernières années a généré une augmentation importante du commerce international (Totonchi et Manshady, 2012). Grâce aux avantages d'Internet, les acquisitions et les ventes sont devenues plus automatisées et plus pratiques, les entreprises sont devenues interconnectées et se sont également connectées avec leurs clients. De plus, le temps et les distances, qui représentaient des barrières commerciales et exigeaient des coûts élevés dans le passé, se sont considérablement raccourcis (Kotler, 2003). En tant que nouvelle méthode commerciale, le commerce électronique a révolutionné le fonctionnement et la gestion des organisations, et présente un énorme potentiel pour les opérations de fabrication, de vente au détail et de services (Gunasekaran et al., 2001).

De nos jours, les consommateurs sont de plus en plus attirés par les achats en ligne en raison du gain de temps, de la flexibilité des prix et de la grande variété de produits que l'on peut trouver sur une plateforme (Ferri et al., 2013). Grâce à la confiance que les entreprises électroniques ont suscitée chez les clients, de nombreux utilisateurs ont choisi de transférer une partie importante de leurs achats vers des plateformes virtuelles (Micu et al., 2017). Cependant, malgré ses nombreux avantages, le commerce électronique peut devenir un gros problème si les entreprises ne sont pas capables de bien gérer leurs opérations. Ainsi, un des problèmes les plus importants dans le commerce électronique de détail est le taux de retour élevé, qui peut atteindre 50% (Kristensen et al., 2013), (Hofacker et Langenberg, 2015).

En raison des pertes importantes que les retours en ligne peuvent générer, les entreprises de commerce électronique utilisent divers outils et technologies pour faire face à ce problème; parmi eux, l’analyse de données et l’intelligence artificielle se démarquent. L’analyse de données et l'intelligence artificielle sont récemment devenues des grands développements technologiques dans le commerce de détail. Ces outils transforment des données vastes et diversifiées en informations précieuses qui peuvent aider à dépasser la vitesse, le coût et la flexibilité vers la chaîne de valeur. En outre, ils peuvent aider à la prévision, au marchandisage et à la planification de leur expansion. (Chitrakorn, 2018.)

En effet, dans la revue de littérature, on trouve des nombreuses études précédentes qui ont été réalisées pour prédire les retours de stock en se concentrant sur la prévision des quantités, en utilisant différentes méthodologies telles que la méthode GARCH (Lu et Perron, 2010), la méthode

ARIMA (Canda et al, 2015), des modèles mixtes avec deux méthodes ou plus (Ma et Kim, 2016), les réseaux de neurones (Kourentzes, 2013), (Enke et Thawornwong, 2005), un algorithme hybride (Zhu et al, 2018), etc. Cependant, connaître le nombre de produits qui seront retournés pourrait ne pas être la meilleure approche pour une entreprise comme notre partenaire, qui fabrique des articles personnalisés, puisqu'il n'y a pas de tailles standards pour chaque type de produit, et chacun de ces produits peut avoir des mesures spécifiques. En ce sens, on considère plus appropriée une approche préventive qui nous permettrait de détecter les personnes qui font le plus de retours et les produits les plus retournés. De sorte qu'il soit possible de prendre des mesures correctives pour éviter ou minimiser le nombre de retours. Ainsi, l'objectif de cette recherche est de développer un modèle capable de trouver le profil de clients qui effectuent les retours, ainsi que les types de produits qui sont retournés.

Pour ce faire, on commencera par trouver des comportements similaires entre nos clients en utilisant la base d'une étude effectuée précédemment pour le développement d'un outil de segmentation des clients (El Ayeb et Agard, 2019). Le résultat nous permettra de reconnaître la variation des mesures morphologiques qu'ont les clients ayant une plus grande tendance à faire des retours. De plus, on utilisera les clusters obtenus comme l'une des variables d'entrée pour la construction d'un arbre de décision, dans le but de trouver les tendances des clients qui effectuent des retours plus souvent. Ensuite, on procédera au développement de notre modèle de prédiction en utilisant les arbres de classification, afin de trouver les caractéristiques des clients qui effectuent des retours. De la même manière, on devra être en mesure de trouver les caractéristiques des vêtements qui sont plus souvent retournés.

La structure de ce mémoire est composée de cinq chapitres. Le premier chapitre nous présente l'introduction, qui nous donne un aperçu du contexte actuel et des problèmes qui ont motivé la conduite de cette recherche. Le deuxième chapitre contient la revue de la littérature utilisée pour soutenir notre recherche. Au cours de l'élaboration de cette section, l'impact du commerce électronique sur la gestion des retours est analysé en profondeur, ainsi que les études précédentes qui ont utilisé différents outils pour résoudre les problèmes et les impacts négatifs de la gestion de retours. Après, on explique ce que sont les séries chronologiques et la nature des séries intermittentes. Ensuite, le sujet des arbres de décision et leur utilisation comme modèles de prédiction sont abordés.

Après avoir clarifié le contexte et la situation du problème en cours d'analyse, au troisième chapitre, on présente la problématique et les objectifs à atteindre, ainsi que la méthodologie de recherche qui est divisée de la manière suivante (1) Nettoyage de la base de données, (2) Analyse statistique préliminaire, (3) Construction de groupes de clients, (4) Construction d'arbres de décision et (5) Analyse des résultats et conclusions. Le quatrième chapitre contient le cas d'étude dans lequel l'application de la méthodologie de ce projet est expliquée étape par étape. Enfin, les conclusions et les recommandations sont présentées au chapitre 5.

CHAPITRE 2 REVUE DE LA LITTÉRATURE

Ce chapitre a pour objectif de donner un aperçu des concepts de base liés avec notre étude. On commence par ce qui concerne l'impact qu'a le commerce électronique a sur la gestion des retours dans les entreprises aujourd'hui. Ensuite, on présente les types de modèles d'apprentissage, ainsi que les algorithmes qui ont été utilisés tout au long du projet. Finalement, la troisième partie va aborder une brève introduction aux séries temporelles.

2.1 Logistique et gestion de retours

La définition de la logistique peut conduire à un débat, car dans la littérature actuelle il existe plusieurs définitions de ce terme qui pointent vers un concept intégré et systématique, elle est fondamentalement orientée vers la satisfaction du client (Ramirez, 2009), elle vise à minimiser les coûts sans réduire la qualité, dans les temps requis, dans la quantité et à l'endroit spécifié dont notre client a besoin.

Le terme logistique est fréquemment associé à la distribution et au transport des produits finis. Il s'agit toutefois d'une appréciation partielle, car la logistique est liée à la gestion des flux de biens et services (Ballou, 2004) depuis l'acquisition de matières premières, et de fournitures à leur point d'origine jusqu'à la livraison du produit fini au point de consommation (Monterroso, 2000).

La logistique peut être définie comme le processus de gestion qui commence par la fourniture des matières premières au point où le produit ou le service est finalement consommé ou utilisé (Lamb et al., 2009), avec trois flux importants: matériaux (inventaires), informations (traçabilité) et fonds de roulement (coûts) (Mora, 2010). D'autres affirment que la logistique est une fonction opérationnelle importante qui comprend toutes les activités nécessaires à l'obtention et à l'administration des matières premières, ainsi que la manutention des produits finis, leur emballage et leur distribution (Ferrell et al., 2010).

De cette manière-là, la logistique n'est pas seulement une fonction de stockage, de manutention et de transport, mais est une méthode de direction et de gestion (Ballesteros et Ballesteros, 2008). En tant que fonction de gestion, la logistique implique des concepts tels que l'emplacement des usines et des entrepôts, les niveaux des stocks, les systèmes d'indicateurs de gestion et d'information, et la gestion des retours (De Brito, 2009).

La gestion d'un système de stocks constitue l'un des aspects logistiques les plus complexes (Zermati et Mocellin, 2005). Les investissements pour la gestion des stocks sont importants, le contrôle du capital associé aux matières premières, aux produits en cours de fabrication et aux produits finaux, constitue un potentiel d'amélioration du système (Gutierrez et Vidal, 2008). Cependant, cette complexité de gestion devient de plus en plus aiguë, compte tenu des effets générés par des phénomènes tels que la mondialisation, l'ouverture des marchés, l'augmentation de la diversification des produits, la production et la distribution de produits avec des normes de qualité élevées ainsi que la massification de l'accès à l'information.

Dans certains secteurs où le taux de retour est très élevé, la gestion des stocks doit faire autant d'effort, voir plus pour la gestion des retours et des commandes afin de répondre correctement à la demande. Lors de la gestion des commandes, il est possible de contrôler l'arrivée des produits et la quantité; cependant, les retours sont plus difficiles à contrôler. Par conséquent, si le niveau de retour est trop élevé, nous devons nous assurer que les articles retournent au processus de vente dès que possible, dans le but d'éviter une surproduction inutile (Hjort, et Lantz, 2016).

Dans quelques entreprises modernes, il est de plus en plus courant de voir comment les produits ou les matériaux sont récupérés auprès des clients (Bruzzoni et al., 2014), (Yuan et al, 2014), soit pour récupérer de la valeur, soit comme services après-vente. Ce processus inverse était appelé il y a des années de la logistique inverse (Luttwak, 1971). La logistique inverse fait partie d'une tendance appelée «la chaîne d'approvisionnement inversée», où les fabricants intelligents conçoivent des processus efficaces pour réutiliser leurs produits (Guide et Van Wassenhove, 2002).

Le processus de logistique inverse est l'ensemble d'activités liées à la planification, la mise en œuvre et le contrôle efficace du flux de matières premières, des stocks actuels, des produits finis et des informations (de Brito et al., 2002), du point de consommation au point d'origine dans le but de les récupérer, de créer de la valeur (Dekker et al., 2013) ou de les jeter (Rogers et Tibben-Lembke, 2003).

Les politiques des entreprises en matière de logistique inverse étaient souvent traitées superficiellement par les gestionnaires. Au fil des années, les pratiques de gestion et les études menées dans ce domaine ont permis d'identifier les attentes les plus importantes de la direction de l'entreprise ainsi que les moyens de mesurer la performance dans ce domaine. Les résultats de la

gestion des retours incluent des facteurs liés aux services clients, mais aussi des facteurs de coûts liés à la récupération de valeur, la rentabilité et les stocks (Jeszka, 2014).

Cependant, aujourd'hui, les technologies de l'information sont indispensables pour les entreprises, et leurs partenaires dans la chaîne d'approvisionnement. Les nouvelles technologies améliorent les performances des entreprises à tous les niveaux (Bharadwaj, 2000). Des études empiriques ont confirmé que l'utilisation des technologies de l'information est cruciale pour la performance de la gestion logistique (Closs et Savitskie, 2003, d'après: Olorunnivo, 2010). Pour cette raison, il a été supposé que les résultats de la gestion des retours et les économies potentielles sont liés au degré d'informatisation de la gestion des retours et de l'utilisation des solutions informatiques par une organisation.

2.2 Commerce électronique

Le commerce électronique désigne l'achat et la vente de biens ou de services en ligne (Rouse, 2012), qui nécessitent l'utilisation de systèmes électroniques tels que l'internet et d'autres réseaux, ainsi que le transfert d'argent et de données pour exécuter ces transactions (Orendorff, 2019). Il comprend également les processus d'échange entre entreprises, les processus internes utilisés pour soutenir leurs achats, ventes, embauches, planification, etc. (Graf et Schnaider, 2016).

Le volume de transactions en ligne a augmenté de façon exponentielle en raison de la pénétration et de la propagation d'Internet. De nos jours, une grande variété de transactions commerciales telles que la gestion de la chaîne d'approvisionnement sont effectuées de cette manière.

Aujourd'hui, le commerce électronique a gagné en popularité, car les technologies utilisées évoluent à pas de géant et sont conviviales et orientées vers le client (Mohapatra, 2013). Ces dernières années ont connu une croissance explosive des détaillants de commerce électronique et de vente au détail, qui devrait atteindre 4 000 milliards de dollars d'ici 2020. De nombreuses études ont aussi montré qu'environ un tiers de toutes les commandes de commerce électronique entraînent des retours chaque année (Zhu, et al., 2018).

La firme américaine spécialisée en marketing eMarketer indique que la part du commerce électronique dans les ventes totales au Canada passerait de 5,9% à 8,2% en 2018 (Bolduc, 2014). De plus, une étude réalisée par Canadapost (2020) nous indique que les dépenses d'achats en ligne se sont élevées à 65 G$ en 2019 et que ces dépenses pourraient atteindre 108 G$ d'ici 2023.

Selon Lalonde (2017) au Québec, les internautes ont dépensé 8,5 milliards de dollars en ligne en 2016, un montant qui représente une augmentation de 6% par rapport à 2015. De plus, une étude réalisée par le CEFRIO (2017) soutient que 57% des adultes québécois ont réalisé des achats en ligne en 2016, comparativement à 58,1% en 2017, soit en augmentation de 1.1%.

2.2.1 Modalités du commerce électronique

Il existe différents types ou catégories de commerce électronique qui sont liés de différentes manières dans un échange commercial. Selon le type de relation qui existe entre l'acheteur et le vendeur, les types de commerce électronique suivants existent:

2.2.1.1 Business to business – B2B (entre entreprises)

Dans ce modèle, l'échange commercial a lieu entre deux entreprises, c'est-à-dire que l'une des sociétés agit en tant que fournisseur pour l'autre, qui adopte le profil du client (Malca, 2001). Les transactions sont généralement liées au commerce de gros, ils ne comprennent pas un seul achat, mais deviennent généralement des relations à long terme, car les entreprises cherchent à établir des relations de confiance avec des fournisseurs qui sont en mesure d'adapter les produits afin qu'ils puissent répondre à leurs besoins commerciaux spécifiques (Carter, 2018).

Le commerce B2B implique l'utilisation de technologies basées sur le Web pour faire des affaires entre entreprises, son aspect essentiel est l'interaction de système à système, ainsi ce type de commerce électronique est moins visible du public. Des entreprises comme Microsoft, IBM, SAP et UPS sont des exemples représentatifs de ce premier modèle.

2.2.1.2 Consommateur à consommateur – C2C (entre clients)

Cette modalité du commerce électronique implique d'effectuer des transactions entre les consommateurs par le biais de processus qui tentent de connecter les fournisseurs privés et les demandeurs sans avoir besoin d'intermédiaires (Reig Torregosa, 2017).

Les stratégies suivies par le commerce C2C sont spécifiées dans trois profils fondamentaux: en tant que forum virtuel dans lequel les relations entre acheteurs et vendeurs privés sont établies, en tant que catalogue illustré à travers lequel effectuer des comparaisons de produits et de prix et en tant que plateforme où les enchères électroniques peuvent être organisées. Parmi les entreprises qui utilisent ce type de modèle figurent E-Bay, Alibaba et Kijiji.

2.2.1.3 Du consommateur aux entreprises – C2B (entre l'entreprise et le client)

Cette variante du commerce électronique est présentée comme une transaction où c'est le client qui impose les conditions de l'échange (Arrechea, 2011). Le client crée de la valeur qui doit ensuite être acceptée, et consommée par l'entreprise (García Gonzales, 2014). En d'autres termes, l'utilisateur utilise les supports numériques à sa disposition pour obtenir des conditions optimales dans l'offre de l'entreprise. On peut voir le modèle C2B fonctionner dans des blogues ou des forums sur Internet dans lesquels l'auteur propose un lien vers une entreprise en ligne, facilitant ainsi l'achat d'un produit (comme un livre sur Amazon.com), pour lequel l'auteur pourrait percevoir un revenu.

2.2.1.4 Entreprise à consommateur – B2C (entre entreprise et client)

Cette modalité entre entreprises et consommateurs, familièrement connue sous le nom d'e-retailing, représente l'application de la culture commerciale traditionnelle au marché virtuel (Li et Li, 2015), constituant ainsi le commerce électronique le plus standardisé (García Gonzales, 2014), (Wang et al., 2020). La relation commerciale entre les entreprises et les consommateurs concerne les processus de commerce électronique entre les entreprises qui ont une certaine offre de biens et services et les individus qui se présentent comme des consommateurs. Le vendeur commence par faire la promotion des produits qu'il offre via son site Web jusqu'à l'achèvement de la transaction initiée en imputant le produit acheté sur la carte de crédit ou de débit du consommateur (O'Conell, 2002). Il prend également en charge le processus de distribution physique et sa surveillance ou la distribution elle-même dans le cas de produits numériques. Amazon est l'une des entreprises les plus populaires utilisant ce modèle.

2.2.2 Dans l'industrie du vêtement

L'essor du commerce électronique a non seulement changé les magasins traditionnels, mais aussi comment et à quelle fréquence les gens font leurs achats, cela vaut particulièrement pour l'industrie de la mode. De nos jours, les consommateurs peuvent profiter de nombreuses plateformes en ligne pour faire livrer des articles à leur domicile et ils peuvent même souvent retourner les pièces sans frais supplémentaires (Simpson et al., 2016). Cela a un impact important sur le comportement des consommateurs, car ils sont encouragés à acheter plus (Janakiraman et al, 2016). Par conséquent, les détaillants concentrent leurs efforts sur l'amélioration de leur réactivité et la réduction extrême des délais de livraison (Simpson et al, 2016), y compris des options de retour faciles dans le cadre

de leur proposition de valeur (Griffis et al., 2012). Mais malgré le confort qu'ils offrent, les commerces électroniques dans l'industrie de la mode ont entraîné des volumes d'expédition élevés pour les livraisons et les retours, qui entraînent en conséquence des coûts élevés.

Les technologies numériques gagnent en importance puisqu'elles fusionnent avec la plupart de nos appareils tels que les téléphones portables et les ordinateurs portables avec de multiples moyens d'accès. De nos jours, il y a des innovations importantes qui poussent le secteur de la mode à être plus efficace, afin d'améliorer l'expérience client et de contribuer à l'amélioration de la proposition de valeur client en général. Les entreprises de mode qui utilisent efficacement les technologies appropriées pourraient accroître leur avantage concurrentiel pour la personnalisation des produits et des expériences d'achat, filtrant le processus logistique qui grignote les budgets (Chitrakorn, 2018).

Le marché mondial de la mode en ligne devrait atteindre 765 milliards de dollars d'ici 2022, soit une augmentation prévue de 281 milliards de dollars, ou 58%, par rapport à l'année 2018. En fait, dans 4 ans, 36% du total des ventes au détail de mode devraient se produire en ligne, contre 27% cette année, selon une récente prévision de Forrester (Johnson, 2018). Pendant l'année 2018, 58% de la population mondiale en ligne a effectué un achat en ligne et environ la moitié d'entre eux ont acheté des vêtements, des accessoires ou des chaussures. Le commerce électronique a un impact plus important sur le marché de la mode, que le commerce de détail au sens large: il a représenté 27% des ventes mondiales de mode en 2018, contre 15% des ventes au détail mondiales (Marketing charts, 2015).

La clé du succès du commerce électronique réside dans la facilité qu'il offre, lors d'un achat ou d'une acquisition, car l'achat peut être reçu directement à domicile dans un délai de plus en plus court. En revanche, l'un des principaux inconvénients pour les détaillants qui vendent en ligne exclusivement est qu'ils peuvent devoir travailler plus dur pour instaurer la confiance et établir une relation avec leurs clients. L'interaction personnelle est limitée lors de la vente en ligne et trouver des clients fidèles peut devenir une tâche difficile.

En outre, étant donné que les clients ne sont pas en mesure de tester ou d'essayer des produits, le commerce électronique a entraîné une augmentation des taux de retours. Ce processus peut coûter cher, et faire reculer les entreprises, si elles ne sont pas prêtes à répondre rapidement aux demandes de leurs clients. Il est important de maintenir la disponibilité des produits, et de s'assurer que le

plus de stock possible peut être revendu au prix fort. Si le processus de retour n'est pas géré avec diligence, le coût des démarques et des ventes perdues continuera d'augmenter.

2.2.3 Gestion de retours

De nombreuses études (Forrester, 2008), (Donaldson, 2015), (Lazar, 2019) ont montré qu'environ un tiers de toutes les commandes de commerce électronique génèrent des retours chaque année. Les coûts de retour directs, tels que les frais d'expédition, de réapprovisionnement, de remise à neuf, ainsi que les coûts indirects, tels que la demande des centres d'appels et la satisfaction des clients, sont devenus un défi majeur pour l'industrie du commerce électronique (Zhu et al., 2018). De plus, lorsqu'un produit est retourné ou échangé, non seulement le détaillant subit des coûts de chaîne d'approvisionnement supplémentaires, mais souvent l'article ne peut pas être revendu au prix d'origine en raison de possibles dommages ou obsolescence compte tenu du temps (mode ou marchandise saisonnière) (Dennis, 2018).

Selon le cofondateur et PDG de Happy Returns, environ 5 à 10% des achats en magasin sont retournés, mais les pourcentages pour les achats en ligne peuvent atteindre de les 40%. Les taux de retour moyens varient selon la catégorie, mais les vêtements et chaussures achetés en ligne affichent généralement les taux les plus élevés, 30 à 40% étant retournés (Reagan, 2019). Par conséquent, les retours deviennent de plus en plus importants à considérer, lors d'un achat en ligne.

Une étude faite par Forrester (2008) a révélé que les retours du commerce électronique étaient la faute du détaillant dans 65% des cas (Lazar, 2019). Les résultats ont révélé que 23% des retours sont dus au mauvais article expédié, 22% des retours sont dus à la différence d'apparence du produit et 20% des retours sont dus à la réception d'un article endommagé.

Pour cette raison, les magasins en ligne sont obligés actuellement d'offrir une expérience de retour appropriée et qui répond aux exigences des clients. Cependant, cette flexibilité a entraîné des pertes importantes pour les entreprises de commerce électronique en raison de la logistique de livraison impliquée et de la valeur de revente éventuellement moindre du produit retourné (Joshi et al., 2018).

Aujourd'hui, les technologies Web et le commerce électronique permettent aux entreprises de gérer les retours plus efficacement en se concentrant principalement sur quatre solutions possibles : minimiser de manière proactive les retours, réduire les facteurs d'incertitude lorsque les retours se produisent réellement, gérer les retours systématiquement par le développement d'applications, et

la réorientation des retours vers de nouvelles destinations grâce à la logistique inverse (Kokkinaki et al., 2002), (Bernon et al, 2016).

En effet, en raison de ces problèmes, l'industrie du commerce électronique élargit progressivement ses horizons avec l'utilisation de différents outils tels que la gestion de connaissances (Shivakumar et Suresh, 2014) et l'intelligence artificielle (Lingam, 2018), qui était initialement utilisée pour prédire le comportement des consommateurs et l'inventaire requis (Chitrakorn, 2018). Les algorithmes d'apprentissage automatique, jouent un rôle crucial dans l'analyse des données concernant les campagnes de marché, et la prévision de l'inventaire (Lingam, 2018) et actuellement, de nombreuses études sont réalisées en utilisant des techniques d'apprentissage automatique pour prévoir les retours de stock (Abe et Nakayama, 2018).

Plusieurs études précédentes se sont intéressées à la prévision de retours de stock dans différents domaines, et à l'aide de différents outils (Chih et al., 2011). (Canda et al., 2015) ont développé un modèle de prévision de retours en utilisant les méthodes ARIMA et Holt. (Ma et Kim, 2016) ont testé un modèle hybride entre ARIMA et DLM. Plus tard (Qi et Zhang, 2008) et (Enke et Thawornwong, 2005) ont exploré l'utilisation des réseaux de neurones, ainsi que d'autres méthodes d'apprentissage automatique, telles que le Random Forest et les machines de renforcement de gradient, afin de prévoir les retours des clients. Cependant, les travaux cités ci-dessus se concentrent sur la prévision des quantités de stock qui seront retournés. Pour une entreprise qui produit des vêtements sur mesure, cette approche n'est pas la plus appropriée, car un seul article pourrait présenter plus de 1000 variations.

D'autre part, (Joshi et al., 2018) ont centré leur recherche sur une entreprise de commerce électronique, dans laquelle la plupart des retours de vêtements étaient placés en raison de la taille. Ils ont donc développé un modèle qui combine la science des réseaux de neurones, et l'apprentissage automatique sur une base des données d'achat et de retour, afin de prédire si un client effectuera un retour ou non. Également, Zhu et al. (2018) ont réalisé l'étude d'un algorithme local de segmentation, basé sur la marche aléatoire pour prédire la probabilité de retour d'un produit pour chaque client, ainsi qu'un graphique hybride pondéré pour représenter les informations riches de l'historique d'achat et de retour des produits, afin de prédire si un client est plus propice à faire des retours. Dans les deux études précédentes, les auteurs ont utilisé des séries temporelles pour pouvoir visualiser le comportement de leurs clients et s'en servir pour obtenir leurs résultats. Ce

type d'approche est plus approprié pour une entreprise comme notre partenaire, car de cette manière les mesures nécessaires peuvent être prises pour éviter de futurs retours.

Finalement, (Sorensen et al., 2000) et (Wang et Chang, 2006) ont utilisé les arbres de classification et régression pour la prévision de retours, car ce sont des modèles flexibles qui peuvent s'adapter au problème et sont plus faciles à interpréter. Ainsi, nous pouvons voir qu'il existe plusieurs solutions qui utilisent différents modèles d'apprentissage et d'analyse de données pour prédire les retours de stock.

2.3 Modèles d'apprentissage

L'apprentissage automatique a connu une croissance et un développement important, il a reçu une grande attention de la part des scientifiques et, pour cette raison, il a participé activement à un certain nombre d'études, contribuant ainsi à résoudre des problèmes réels ayant un grand impact sur la société. Les experts définissent l'apprentissage automatique, comme la branche de l'intelligence artificielle dédiée à l'étude de programmes d'apprentissage basés sur leur expérience, pour effectuer de mieux en mieux une tâche donnée. L'objectif principal d'un modèle d'apprentissage automatique est d'utiliser les preuves, ou informations fournies comme données d'entrée à titre d'exemples, afin de créer une hypothèse et de pouvoir répondre à une situation inconnue (Mitchell, 1997). Les experts suggèrent différentes approches de l'apprentissage automatique, notamment : apprentissage supervisé, apprentissage non supervisé et apprentissage par renforcement.

2.3.1 Modèles d'apprentissage supervisé

L'apprentissage supervisé est le modèle d'apprentissage le plus courant utilisé pour la résolution de problèmes. On dit qu'un problème peut être résolu avec un algorithme d'apprentissage supervisé, lorsque chaque exemple de données (normalement présenté comme un vecteur d'attributs X) est étiqueté à une classe d'appartenance prédéfinie Y. Pour cette raison, la tâche principale est de concevoir un algorithme automatique qui apprend une règle de classification dans un ensemble de données appelées données d'apprentissage, qui produit ensuite la sortie idéale pour chaque exemple des données de test. En d'autres termes, l'algorithme apprend des règles de décision à partir de données connues pour prédire l'étiquette des données inconnues. Dans ce groupe d'algorithmes, on peut distinguer une subdivision en fonction du type d'étiquette à prédire.

2.3.1.1 Arbres de Classification

Les arbres de décision sont des modèles de prédiction dont l'objectif principal est l'apprentissage inductif à partir d'observations et de constructions logiques (Gehrke et al., 1999). Ils sont très similaires aux systèmes de prédiction basés sur des règles, qui servent à représenter et à catégoriser une série de conditions, qui se produisent successivement pour résoudre un problème. Ils sont probablement le modèle de classification le plus connu et le plus utilisé (Quinlan, 1986).

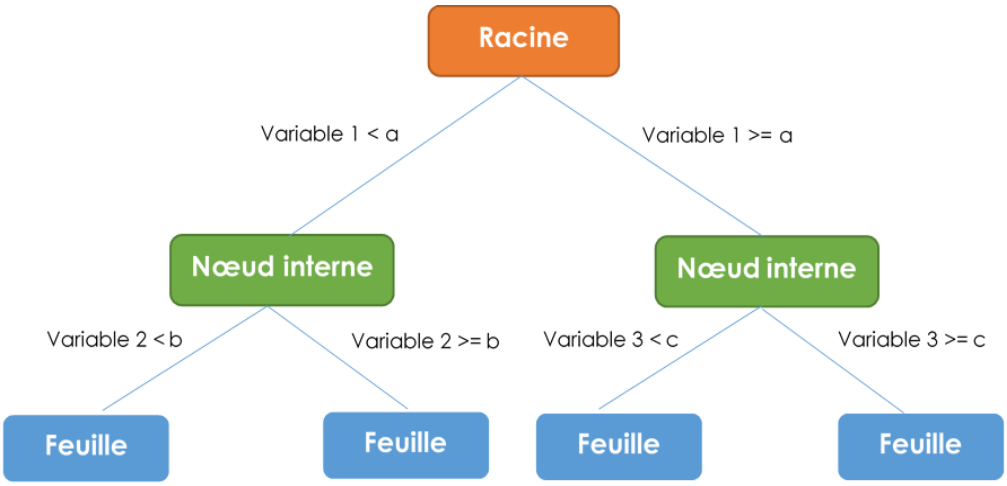


Figure 2.1 Exemple de division d'un arbre de classification

Les connaissances acquises au cours d'un processus d'apprentissage inductif sont représentées par un arbre. Cet arbre (Figure 2.1) est représenté graphiquement par un ensemble de nœuds, de feuilles et de branches. Le nœud principal ou racine est l'attribut à partir duquel le processus de classification commence; les nœuds internes correspondent à chacune des questions sur l'attribut particulier du problème. Chaque réponse possible aux questions est représentée par un nœud interne. Les branches qui sortent de chacun de ces nœuds sont étiquetées avec les valeurs possibles de l'attribut. Les nœuds finaux, aussi appelés feuilles correspondent à une décision qui coïncide avec l'une des variables de classe du problème à résoudre.

Ce modèle est construit à partir de la description narrative d'un problème, car il fournit une vision graphique de la prise de décision, en spécifiant les variables qui sont évaluées, les actions qui doivent être prises et l'ordre dans lequel la prise de décision sera effectuée. Chaque fois que ce type de modèle est exécuté, un seul chemin sera suivi en fonction de la valeur actuelle de la variable évaluée. Les valeurs que les variables peuvent prendre pour ce type de modèle peuvent être discrètes ou continues.

La plupart des algorithmes utilisés pour construire un arbre sont des variantes d'un algorithme générique, appelé « algorithme gourmand » qui va essentiellement de la racine vers le bas (de haut en bas) à la recherche récursive des attributs, qui génèrent le meilleur arbre jusqu'à trouver l'optimum global avec une arborescence aussi simple que possible (Salzberg, 1994). Les algorithmes les plus connus sont ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993), C5.0 (Kuhn et Johnson, 2013) et CART (Classification And Regression Trees) (Breiman et al, 1984). La différence entre les 3 premiers et CART, est la possibilité pour les derniers d'obtenir des valeurs réelles, en tant que résultats par rapport à des valeurs discrètes dans le reste des modèles (Patil et al., 2012). Les principales caractéristiques de ces algorithmes, sont leur capacité à traiter efficacement un grand volume d'informations, et leur capacité à gérer le bruit (erreur dans les valeurs ou leur classification) qui peut exister dans les données d'apprentissage. En général, les différents modèles d'apprentissage diffèrent par leur algorithme qui classe les différents attributs de l'arbre et leur efficacité afin d'obtenir un meilleur arbre aussi simple que possible.

Les arbres de classification présentent une méthode qui est assez similaire à celle des arbres de régression. La différence est que sa variable de sortie est de type catégorie. La façon de créer des partitions utilise la division binaire récursive. Cependant, contrairement aux arbres de régression, nous ne cherchons pas à minimiser la somme résiduelle des carrés, mais plutôt le taux d'erreur de classification, cette erreur est donnée par l'équation suivante (2.1) :

$$E = 1 - max_k(p_{mk}) \tag{2.1}$$

Où p_{mk} représente la proportion des observations de formation dans la m-ième région qui sont de classe k.

Les méthodes fondamentales en C5.0

Comme ses prédécesseurs, cet algorithme construit les arbres à partir d'un ensemble de données d'apprentissage optimisé sous les critères d'entropie et de gain d'information et correspond à une évolution de sa version précédente, l'algorithme C4.5 (Quinlan, 1993). Parmi ses caractéristiques, la capacité à générer des arbres de prédiction simples, des modèles basés sur des règles, des ensembles basés sur l'augmentation, et l'attribution de différents poids aux erreurs se démarquent (Kuhn et al, 2015).

Le modèle divise l'échantillon en fonction du champ qui offre le gain d'information maximal à chaque niveau. Le champ cible doit être catégorique. Diverses divisions sont autorisées dans plus de deux sous-groupes (Vallejo et Tenalanda, 2012). Cet algorithme a été très utile pour créer des modèles de classification et toutes leurs capacités en r en utilisant le package C50. Les plus grands avantages de cette version ont à voir avec l'efficacité dans le temps de construction de l'arbre, l'utilisation de la mémoire et l'obtention d'arbres considérablement plus petits que dans le C4.5 avec la même capacité prédictive (Weiss et al., 2007).

L'entropie

Pour comprendre le fonctionnement de cet algorithme, il est nécessaire de comprendre le terme de pureté. Si les segments de données contiennent des éléments d'une seule classe, ils sont considérés comme purs, sinon les segments sont impurs. L'entropie est la mesure de l'impureté utilisée par l'algorithme C5.0, et elle est définie par l'équation 2.2 :

$$Entropie(S) = -\sum_{i=1}^{c} p_i \, log_2(p_i) \qquad (2.2)$$

Où :

c : le nombre total de niveaux de classe e

p_i : La proportion d'observations qui entrent dans chaque classe (c'est-à-dire la probabilité qu'un point de données sélectionné au hasard appartienne au i-ème niveau de classe).

Pour deux classes possibles, l'entropie va de 0 à 1. Pour n classes, l'entropie va de 0 à $\log_2$ (n), où l'entropie minimale correspond à des données purement homogènes et l'entropie maximale représente des données complètement désordonnées. Plus l'entropie est petite, plus il y a d'informations contenues dans la division des nœuds de décision correspondante. Les données avec une entropie élevée indiquent un contenu d'information très aléatoire, et les données à faible entropie indiquent des données hautement compressibles avec une structure.

Gain

Le gain d'information, définie par l'équation 2.3, est l'indicateur qui mesure la qualité d'une variable. Il est basé sur la diminution de l'entropie après la division d'un ensemble de données en attribut.

$$\text{Gain (F)} = Entropie(S) - Entropie(S_1) \qquad (2.3)$$

Les arbres obtenus dans ce travail sont modélisés avec l'algorithme C5.0, afin d'utiliser une grande quantité d'informations et de mieux gérer les variables continues. De plus, les arbres obtenus sont plus simples et plus faciles à comprendre.

2.3.2 Modèles d'apprentissage non supervisé

Contrairement à l'apprentissage supervisé, dans ce type d'apprentissage, l'attribut dépendant n'est pas connu « a priori », c'est-à-dire qu'il n'y a pas de classification des exemples en tant que tels, l'apprentissage est guidé par la similitude et la dissemblance entre les exemples donnés (Giner et al., 2016). Les principales applications de l'apprentissage non supervisé sont : la segmentation des ensembles de données, la détection d'anomalies et la simplification des ensembles de données en ajoutant des variables avec des attributs similaires.

2.3.2.1 Segmentation

Le processus de regroupement d'un ensemble d'objets, physiques ou abstraits en classes d'objets similaires est appelé segmentation. L'objectif de la segmentation est de diviser un ensemble de données en groupes d'instances de données ou d'objets en utilisant des mesures de distance ou probabilistes (Makhabel, 2015). Un cluster est défini comme une collection d'objets de données, présentant des similitudes entre eux au sein du même cluster, et qui possèdent des caractéristiques différentes des objets d'autres clusters (Han et al., 2011). Si on considère une perspective d'apprentissage automatique, les clusters correspondent à des modèles cachés, la recherche de clusters est un apprentissage non supervisé, et le système résultant représente un concept de données (Berkhin, 2006).

Les algorithmes d'apprentissage non supervisés ont une gamme incroyablement large d'applications, et sont très utiles pour résoudre des problèmes du monde réel tels que la détection d'anomalies, les systèmes de recommandation, le regroupement de documents ou la recherche de clients ayant des intérêts communs en fonction de leurs achats. Afin d'améliorer leur performance, on a besoin généralement d'un large ensemble de modèles de formation qui seront utilisés par le classificateur pour modéliser chacun des groupes.

- **Méthodes hiérarchiques**

  Les méthodes hiérarchiques construisent une hiérarchie de cluster ou, en d'autres termes, un arbre de clusters, également appelé dendrogramme (Verma et al., 2012). Ils sont basés sur un processus séquentiel de formation de groupes ou clusters, en établissant des hiérarchies entre objets ou individus. Les méthodes hiérarchiques sont divisées en méthodes par agglomération et par division (Kaufman et Rousseeuw, 2009). Les méthodes par agglomération commencent par un nombre de groupes égal au nombre total d'objets et ils sont joints de manière « agglomérative » selon une métrique de distance; à la fin, un seul cluster est formé avec tous les objets. D'un autre côté, le processus des méthodes de division commence par un seul groupe ou cluster, qui contient tous les objets. Ensuite, ils sont divisés (par ordre décroissant) en sous-groupes en considérant les plus éloignés, jusqu'à atteindre « n » groupes. Le plus utilisé est: l'algorithme de Howard-Harris.

- **Méthodes de partitionnement**

  Une méthode de partitionnement crée k groupes à partir de n objets non étiquetés de la manière dont chaque groupe contient au moins un objet (Leung et al., 2000). L'un des algorithmes les plus utilisés de partitionnement est k-Means (Forgy, 1965), où chaque cluster a un prototype qui est la valeur moyenne de ses objets. L'idée principale derrière k-Means est la minimisation de la distance totale (généralement la distance euclidienne) entre tous les objets dans un cluster de leur centre de cluster (prototype).

- **Méthodes basées sur la densité**

  Ces méthodes sont basées sur la détection des zones dans lesquelles il y a des concentrations de points et où elles sont séparées par des zones vides ou avec peu de points (Ware et Bharati, 2013). Les points qui ne font pas partie d'un cluster sont étiquetés comme du bruit. Cette méthode peut être utilisée pour filtrer le bruit (valeurs aberrantes) et découvrir des groupes de formes arbitraires.

Il existe également d'autres méthodes de segmentation telles que les méthodes par grilles et les méthodes basées sur des modèles.

2.3.2.2 Dynamic Time warping

Le Dynamic Time Warping (DTW) trouve son origine dans le domaine de la reconnaissance vocale (Rabiner et Juang, 1993), (Adams et al., 2005), où elle a montré une bonne performance. Ses qualités en termes de reconnaissance permettent un alignement optimal entre deux séquences vectorielles de longueurs différentes, contrairement à la distance euclidienne conventionnelle (Aghabozorgi et al., 2015) en utilisant la programmation dynamique. Comme résultat, une mesure de distance entre les deux séries de temps est obtenue (Sato et Kogure, 1982). Cet algorithme a trouvé des applications dans diverses disciplines telles que: l'exploration de données, la reconnaissance des gestes, la robotique, les processus de fabrication ou la médecine (Keogh et Pazzani, 2001). Dans l'exploration de données de séries temporelles, le DTW est couramment utilisé pour calculer la différence entre deux séries (Liu et al., 2002), (Keogh et Pazzani, 2001).

L'algorithme DTW commence par créer une matrice de distance entre le patron et les points d'échantillonnage, puis il trouve la meilleure façon d'aligner l'échantillon avec le patron, en recherchant le chemin optimal (Van der Vlist et al., 2019), (Wang et Gesser, 1997).

Il existe certaines restrictions, locales et globales, avec lesquelles on essaie d'améliorer l'algorithme, soit en minimisant la distance, soit en réduisant le nombre de calculs. Les restrictions locales recherchent le chemin avec la distance la plus courte en faisant varier la pente (Zoltan et al, 2015). Quant aux limites globales, elles consistent à faire varier la fenêtre, sa taille ou sa forme, pour essayer de minimiser le nombre de points à calculer. Les deux types de fenêtres par excellence sont le parallélogramme (Itakura, 1975) et la bande (Sakoe et Chiba, 1978).

Si on a deux séries de nombres réels A et B, de taille I et J respectivement :

$$A = a_1, a_2, .., a_i, \dots, a_I$$

$$B = b_1, b_2, .., b_j, \dots, b_J$$

Et on considère dans le plan des axes i,j, les alignements entre les indices de deux séries, à partir des trajectoires F à calculer sous la forme :

$$F = c_1, c_2, \dots, c_k, \dots, c_K$$

Où : $c_k = (i_k, j_k)$ sont des points dans le plan

Avec

$$\max(I, J) \leq K \leq I + J \tag{2.4}$$

Pour chaque point c_k :

$$d_k(c_k) = d_k(i_k, j_k) = ||a_{ik} - b_{ik}||$$

d_k représente la distance entre les deux valeurs des deux séries alignées. Étant donné une trajectoire F, la somme pondérée de ces distances est définie par l'équation 2.5 :

$$S(F) = \sum_{k=1}^{k=K} d(c_k) w_k \quad , w_k \geq 0 \tag{2.5}$$

Lorsque toutes les différences dues aux "déformations" dans le temps ont été éliminées, il est raisonnable de penser que l'on trouve le chemin optimal F et que la somme résultante S optimale est la "vraie" distance entre les deux séries. Dans ce sens, on peut définir une mesure DTW de similitude entre deux séries A et B avec l'équation suivante (Fu et al., 2017):

$$DTW(A, B) = min\{S(F)\} \tag{2.6}$$

Ces trajectoires F doivent respecter les restrictions suivantes (Yu et al., 2011), (Lou et Ao, 2015), (Rakthanmanon et al., 2013):

- **Conditions de monotonie**

$$i_{k-1} \leq i_k \; et \; j_{k-1} \leq j_k \tag{2.7}$$

- **Conditions de continuité**

$$i_k - i_{k-1} \leq 1 \; et \; j_k - j_{k-1} \leq 1 \tag{2.8}$$

En conséquence, c_{k-1} peut seulement prendre trois valeurs :

$$c_k = \begin{Bmatrix} (i_k, j_k + 1 \\ (i_k + 1, j_k + 1 \\ (i_k + 1, j_k \end{Bmatrix} \tag{2.9}$$

Ces trois cas correspondent respectivement à une avance sur l'axe des x sur la matrice de coût, une avance sur les axes x et y sur la matrice de coût, ou bien, une avance sur l'axe des y sur la matrice de coût.

Dans tous les cas, nous veillons à avancer sur l'un des axes

- **Conditions de bord**

$$i_1 = 1, j_1 = 1, i_k = I, j_k = J \tag{2.10}$$

Grâce à ces restrictions, nous établissons les limites de la matrice de coûts.

- **Conditions de la pente**

  Afin d'éviter qu'une section courte d'une des séries corresponde à une section longue de l'autre série, la condition suivante est établie, si un point ck se déplace m fois consécutivement sur l'un des axes, il doit se déplacer, au moins n fois dans le sens diagonal, avant de repartir dans le même sens. La façon de mesurer cette restriction est le paramètre : p = n/m. Plus p est grand, plus la restriction est grande. En particulier, si p = 0, il n'y a pas de restriction

Cependant, l'un des principaux inconvénients lors de la mesure de la distance entre deux séries temporelles avec DTW est la vulnérabilité aux interférences dues au bruit et aux valeurs aberrantes de la série chronologique (Wang et al., 2014). Puisque l'algorithme DTW compare les séries point par point, lorsque la série chronologique est trop longue, le calcul peut être très volumineux (Lou et al., 2015).

2.3.2.2.1 Algorithme des pics de densité

L'analyse des clusters vise à classer les éléments en catégories sur la base de leur similitude. L'algorithme basé sur les pics de densité (DP) repose sur l'idée que les centres de cluster ont une densité plus élevée que leurs voisins et qu'ils sont également situés relativement loin des autres points avec les densités les plus élevées (Sieranoja et Fränti, 2019). Au cours de la procédure de segmentation, le nombre de clusters apparaît intuitivement, les valeurs aberrantes sont automatiquement repérées et exclues de l'analyse, et les clusters sont reconnus indépendamment de leur forme ou dimensionnalité, l'une des applications les plus notables est développé dans les travaux de recherche de (Rodriguez et Laio, 2014). En ce sens, l'algorithme DP est capable de gérer des clusters avec des formes arbitraires, ce qui est particulièrement important pour le DTW, car nous ne pouvons pas visualiser les clusters DTW dans un espace métrique (Begum et al., 2015). La figure 2.2 nous montre un exemple de l'algorithme, où on peut apprécier des points en couleur rouge avec la densité la plus élevée qui représentent les centres de chaque cluster.

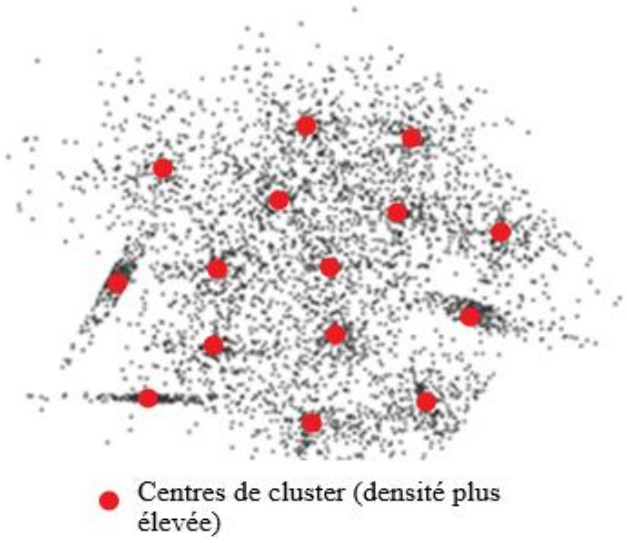


Figure 2.2 Exemple de regroupement des pics de densité

2.4 Séries temporelles

Comme mentionné dans la section précédente, le DTW est utilisé pour trouver des comportements similaires entre séries temporelles, dans la présente étude nous nous sommes confrontés à un type particulier de séries que nous présenterons ci-dessous.

Dans certains domaines scientifiques, on travaille avec des modèles dont la dépendance est considérée comme aléatoire, car les observations varient individuellement (Gimeno et al., 2014). Cependant, dans d'autres domaines, les observations dépendent les unes des autres et sont également faites sur une certaine période, de sorte que, dans une certaine mesure, on peut également parler de dépendance temporelle (Rice, 1963), (Gudmunsson, 1971), (Liebert et Schuster, 1989). Ces séquences de données numériques référencées à la même variable et correspondant à différents moments du temps sont appelées des séries temporelles (Carrión García, 2001). Une série temporelle est définie comme une séquence de N observations ou données équidistantes ordonnées selon une ou plusieurs caractéristiques quantitatives d'un phénomène individuel à différents moments dans le temps, dans lesquels les observations sont faites. Une série temporelle (ou simplement une série) est une séquence de N observations (données) ordonnées (Fu, 2011. Les données peuvent se comporter de différentes manières dans le temps, c'est-à-dire qu'elles présentent une tendance, un cycle; ne pas avoir une forme définie ou aléatoire, des variations saisonnières, qui peuvent être annuelles, mensuelles, etc (Kirchgässner et al., 2007).

Les méthodes traditionnelles pour l'analyse des séries temporelles sont faites par la décomposition de la même série en plusieurs parties (Maindonald, 2009). On dit qu'une série chronologique peut être décomposée en trois composantes qui ne sont pas directement observables, à partir desquelles seules des estimations peuvent être obtenues. L'ensemble des techniques visant à l'analyse de ces séries d'observations dépendantes du temps est appelé l'analyse des séries temporelles (Yafee et McGee, 2000).

2.4.1 Composants d'une série temporelle

L'analyse classique des séries chronologiques est basée sur l'hypothèse que les valeurs prises par la variable d'observation sont la conséquence de trois composantes, dont l'action conjointe aboutit aux valeurs mesurées, ces composantes sont:

- **Composante de tendance**

  Elle peut être définie comme un changement à long terme qui se produit dans la relation au niveau moyen, ou le changement à long terme de la moyenne. La tendance est identifiée avec un mouvement à long terme en douceur dans la série.

- **Composante saisonnière**

  De nombreuses séries chronologiques ont une certaine périodicité ou, en d'autres termes, une variation sur une certaine période (semestrielle, mensuelle, etc.).

- **Composante aléatoire**

  La composante aléatoire mesure la variabilité de la série chronologique après avoir retiré les autres composantes (Cryer, 2008). Cette composante ne répond à aucun modèle de comportement, mais est le résultat de facteurs aléatoires qui ont un impact isolé sur une série chronologique. En ce sens, on peut dire que cette composante se produit dans toutes les séries temporelles et fait référence aux changements qui se produisent dans les valeurs d'une série temporelle en raison de phénomènes qui ne peuvent pas être expliqués, par conséquent, leur occurrence est attribuée au hasard. Par exemple, les données économiques sont parfois considérées comme affectées par des cycles économiques avec une période allant d'environ 3 ou 4 ans à plus de 10 ans selon la variable mesurée (Chatfield, 2019).

2.4.2 Classification des séries temporelles

Il est possible de qualifier les séries chronologiques en tenant compte de leur saisonnalité (Thompson, 1994). L'une des classifications divise ces séries en deux types, stationnaires et non stationnaires. Une série est stationnaire lorsqu'elle est stable dans le temps, c'est-à-dire lorsque la moyenne et la variance sont constantes dans le temps (Escudero et Vallejo, 2000). Cela se traduit graphiquement par le fait que les valeurs des séries ont tendance à osciller autour d'une moyenne constante et que la variabilité par rapport à cette moyenne reste également constante dans le temps. Tandis que les séries non stationnaires présentent une tendance et/ou une variabilité qui évolue dans le temps. Les variations de la moyenne déterminent une tendance à long terme à augmenter ou à diminuer, de sorte que la série n'oscille pas autour d'une valeur constante (Cendejas, 2016).

2.4.3 Séries temporelles intermittentes

Les séries temporelles intermittentes sont des séries, ayant peu d'observations, qui diffèrent grandement d'une période à l'autre, avec des intervalles d'occurrence irréguliers (Croston, 1972). Une demande intermittente ou sporadique survient lorsqu'un produit subit plusieurs périodes sans consommation, c'est-à-dire où la valeur de la demande est égale à zéro (Carrion García, 2001), comme nous pouvons l'apprécier sur la figure 2.3. Souvent, dans ces situations, lorsque la demande se produit, elle est faible et parfois de taille très variable. Les séries temporelles sont un type de données définies comme étant une séquence de valeurs obtenues par des mesures répétitives au cours du temps d'un phénomène précis (Han et al., 2011). (Yang et Wu, 2006) incluent l'analyse des séries temporelles parmi les dix défis de l'analyse de données.

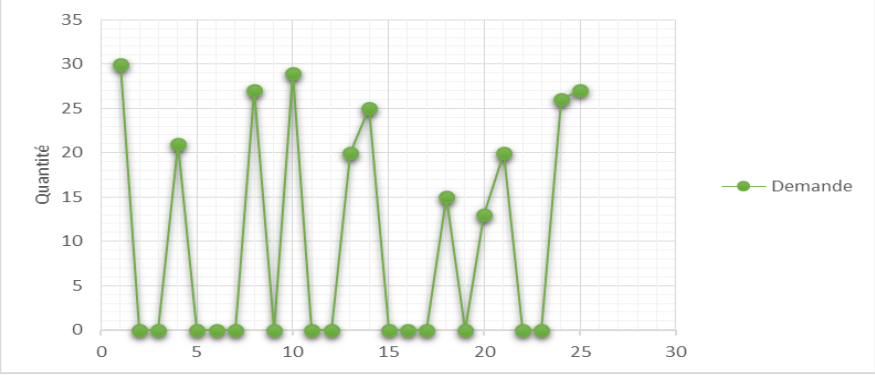


Figure 2.3 Exemple de demande intermittente

Dans la littérature, on trouve plusieurs méthodes qui nous permettent d'agréger les données intermittentes afin de corriger les irrégularités et de rendre possibles leur exploitation. Parmi ces méthodes, on trouve le calcul des moyennes mobiles simples (MMS), le lissage exponentiel simple (SES) et la méthode de Croston (1972) qui est une dérivée du lissage exponentiel. Par exemple, (Murray et al, 2018) ont mis en place un modèle qui se base sur une suite d'agrégation au niveau temporel le plus fin, un lissage en utilisant la méthode de Croston, et finalement une réagrégation temporel.

2.5 Conclusion

Au cours de ce chapitre, nous avons présenté les notions de base pour notre recherche. On a commencé par présenter l'impact du commerce électronique dans l'industrie du vêtement et la gestion de retours. Ainsi que les outils utilisés par les entreprises d'aujourd'hui pour faire face à ce problème. Nous avons procédé à la présentation des arbres de décision qui sont une méthode d'apprentissage supervisé. L'algorithme utilisé pour ce travail est le C5.0, qui permet la construction d'arbres de classification de manière simple, rapide et facile à comprendre. Ensuite, nous avons abordé les algorithmes d'apprentissage supervisé en mettant l'accent sur la segmentation, puisque cette technique sera utilisée pour trouver des comportements similaires, parmi les individus de notre cas d'étude. Finalement, nous avons exposé les séries chronologiques qui constituent un élément fondamental de cette recherche, car leur construction est nécessaire avant de procéder à faire notre segmentation.

CHAPITRE 3 QUESTION DE RECHERCHE ET MÉTHODOLOGIE

3.1 Contexte et problématique

La popularité du commerce électronique a augmenté ces dernières années grâce à la facilité d'accès et à l'efficacité de la distribution, qui nous permet de recevoir un article en 24 heures ou moins. De plus en plus de personnes préfèrent acheter leurs produits en ligne. Bien que ce type de vente présente de nombreux avantages, nous devons garder à l'esprit que l'un des principaux inconvénients, les retours de produits, peuvent entraîner des pertes importantes s'ils ne sont pas gérés correctement. Les retours dans le commerce électronique restent l'un des obstacles les plus importants, puisque les utilisateurs n'ont pas la possibilité de toucher et de tester le produit. Il faut aussi prendre en compte le type de produit offert, car retourner une chemise n'est pas du tout pareil que retourner un meuble.

Dans notre recherche nous nous intéressons plus particulièrement à l'industrie du vêtement. Ce secteur possède le taux de retour en ligne le plus élevé, principalement en raison des facilités qu'elle offre et des problèmes pour trouver la bonne taille. Par conséquent, il est important pour les entreprises d'avoir une bonne stratégie pour la gestion de ses retours, en particulier celles qui ont un volume de commandes important, puisque les retours peuvent coûter deux fois plus cher qu'une livraison. Cette situation a conduit les entreprises à s'adapter.

De nos jours, il existe différentes stratégies et outils utilisés par le commerce électronique pour gérer les retours, y compris l'apprentissage automatique. Les modèles et algorithmes de prédiction sont de plus en plus utilisés car grâce à eux, il est possible de connaître la quantité des retours et de prendre les meilleures décisions en conséquence. C'est dans ce contexte que se situe notre travail.

Comme mentionné ci-dessus, bien qu'il soit possible de prédire le nombre de produits qui seront retournés, cette approche n'est pas considérée comme la plus appropriée pour notre recherche, car notre partenaire fournit des articles personnalisés. Pour cette raison, on a décidé de réaliser un modèle de prédiction qui nous permet de trouver les clients ayant la plus grande tendance à effectuer des retours et les produits les plus retournés. Ainsi, notre partenaire pourra prendre les mesures nécessaires pour éviter de futurs retours.

3.2 Objectifs

3.2.1 Objectif général

L'objectif principal est de développer un modèle qui nous permette de reconnaître les types de clients qui font souvent des retours, et les caractéristiques des produits qui ont tendance à être retournés. Pour atteindre cet objectif, nous procéderons à la recherche de tendances dans l'évolution des mensurations des clients. Ces tendances seront utilisées comme variables d'entrée explicatives. Ensuite, à l'aide d'arbres de décision, nous procéderons à la construction d'un modèle de prédiction.

3.2.2 Objectifs spécifiques

Les objectifs spécifiques de ce travail visent, premièrement à analyser et préparer les données collectées afin qu'elles soient adaptées aux outils d'analyse subséquente. Notamment, nous transformerons nos données en séries temporelles, et réaliserons différentes agrégations temporelles.

Le deuxième objectif est d'identifier les profils de mensurations qui sont similaires. Pour cela un algorithme de segmentation sera utilisé sur les historiques de mensurations des utilisateurs.

Le troisième objectif est de choisir et, de tester un modèle de prédiction et d'en interpréter les résultats.

3.3 Méthodologie

Pour la présente enquête, les informations relatives à l'historique d'achat, et de retour d'un groupe de clients sur une période de six ans sont disponibles. Ces données incluent des informations pertinentes et spécifiques sur chaque client, telles que son identifiant, son sexe et son poste. De plus, les caractéristiques des vêtements commandées par chaque individu sont également connues.

Afin d'exploiter correctement la base de données, il est nécessaire de la préparer au préalable. Ainsi, on commence par préparer et nettoyer la base de données afin de pouvoir effectuer une série d'analyses statistiques, en se concentrant sur les retours de vêtements. On continue à construire des séries temporelles avec l'historique personnel de chaque client, afin d'arriver à visualiser la variation de leurs mensurations tout au long de la période d'étude. Ces séries temporelles nous

permettront de regrouper les clients, en fonction de l'évolution de leur taille dans le temps. Plus tard, les groupes obtenus seront ensuite utilisés comme l'une des variables d'entrée pour nos modèles de prédiction. Nous allons procéder au choix des variables d'entrée pour construire nos arbres de décision, et les retours seront utilisés comme variable de sortie. La figure 3.1 présente la méthodologie développée dans le cadre de ce mémoire.

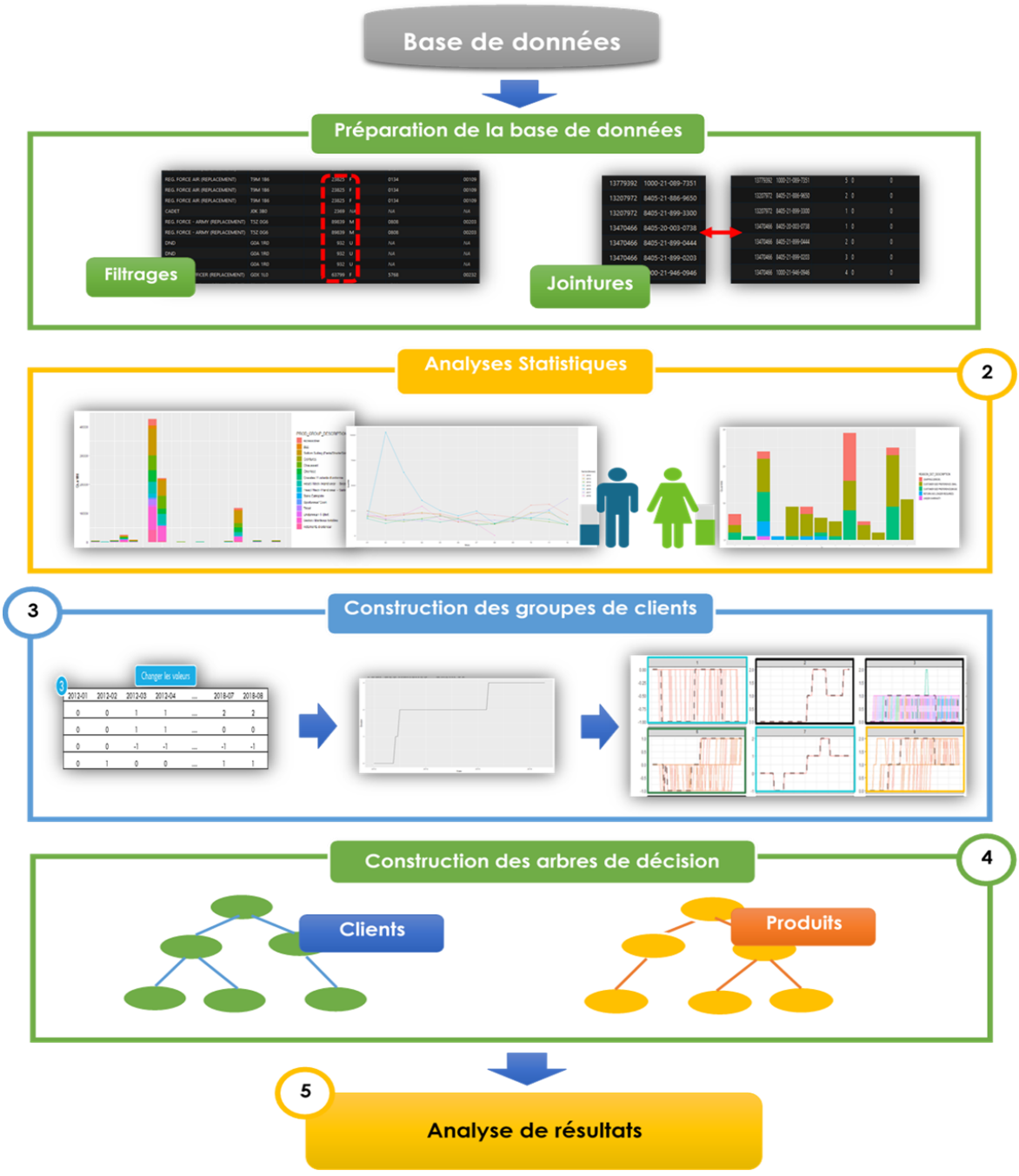


Figure 3.1 Étapes de la méthodologie proposée

Comme on peut le voir sur la figure 3.1, ce travail peut être divisé en 5 étapes principales. (1) La première étape consiste à préparer la base de données (section 3.3.1). Ensuite, on effectue des analyses statistiques (2) pour mieux comprendre la situation actuelle de notre partenaire. La troisième étape consiste en la (3) construction de groupes de clients, où il sera d'abord nécessaire de réaliser des séries chronologiques permettant de visualiser l'évolution des mensurations corporelles des individus, afin de pouvoir regrouper plus tard des tendances similaires par segmentation. Par la suite (4) on va sélectionner les variables d'entrée pertinentes, parmi lesquelles on inclura les clusters de l'étape précédente. Pour plus de précision, les résultats seront comparés à l'aide de différents algorithmes. Finalement (5) la dernière étape consistera à analyser et interpréter les résultats obtenus. On détaillera chacune de ces étapes dans les sections suivantes.

3.3.1 Préparation de la base de données

Lorsqu'on commence à travailler sur un problème d'exploration de données, toutes les informations collectées au sein d'une entreprise industrielle ne peuvent pas toujours être utilisées, sans les avoir d'abord préparées (Liu et Motoda, 2012). Les données doivent être assemblées, intégrées et nettoyées (Witten et al., 2017). La préparation des données comprend des tâches telles que : l'audit des données pour trouver des écarts, le choix des transformations que nous allons corriger et l'application des transformations sur l'ensemble de données (Raman, et Hellerstein, 2001). De ce fait, un traitement correct des données est fondamental dans le développement du projet, car la qualité des résultats qui seront obtenus en dépend.

Cette section comprend les activités générales de traitement et de nettoyage des données, ainsi que l'intégration de différentes bases de données et les changements de format qui seront appliqués. Elle est aussi étroitement liée aux phases 3.3.2 et 3.3.3, car selon la méthode ou l'algorithme à utiliser, les données doivent être traitées de différentes manières. Par conséquent, il convient de noter que des préparations supplémentaires peuvent être nécessaires pour ces deux étapes.

Pour cette étude, notre partenaire industriel nous a fourni différentes bases de données avec des informations liées à l'historique d'achat d'un de ses principaux clients, où l'on trouve plus d'un million de commandes. Chaque base de données contient des informations pertinentes qui peuvent appartenir aux clients tels que leur identifiant, leur sexe, leur fonction et leur emplacement; ou qui peuvent être liées aux produits tels que le type de vêtement et leurs caractéristiques respectives. Cependant, les données n'ont pas été préparées, ou ne sont pas adéquates pour atteindre un objectif

spécifique. Ainsi, cette première étape consiste à étudier la base de données afin de mieux comprendre chaque attribut, sa signification et ses différentes valeurs. De telle manière, qu'on puisse procéder à faire des jointures entre les bases et des filtrages d'information, c'est-à-dire éliminer les données non nécessaires.

La préparation des données étant très spécifique au cas d'étude, elle sera présentée en section 4.2.

3.3.2 Analyse statistique préliminaire

L'analyse statistique des données est un processus très important, qui nous permet d'interpréter les données dont on dispose pour prendre les meilleures décisions (Pearson et Wegener, 2013). Au cours de cette étape, nous examinerons attentivement les données fournies par notre partenaire afin de comprendre le contexte actuel des retours; qui ont été effectués par l'un de ses principaux clients institutionnels.

On dispose des données qualitatives et quantitatives, qui nous permettront de décrire les variables, et d'expliquer les relations possibles entre elles. En outre, cette analyse nous permettra également d'explorer les données, de manière à pouvoir extraire des informations supplémentaires qui peuvent ne pas être évidentes, telle que la distribution des données, la quantité des retours effectués chaque année, les raisons des retours et la découverte de valeurs aberrantes (valeurs extrêmement faibles ou élevées). Avoir ces informations sera très utile pour classer les données et rechercher d'éventuelles erreurs, de même que pour le paramétrage des différents algorithmes d'apprentissage.

Pour représenter les résultats obtenus de manière compréhensible, nous utiliserons des graphiques et des tableaux qui nous permettront d'émettre des conclusions et de prendre des décisions pour les étapes suivantes.

L'analyse des données du partenaire industriel sera présentée en section 4.3.

3.3.3 Construction des groupes de clients

Cette étape vise à diviser les clients en groupes ayant des comportements similaires en fonction de l'évolution de leurs mensurations. De plus, grâce à cette segmentation, nous obtiendrons également une variable supplémentaire pour la construction de notre modèle de prédiction. Pour cela, il faudra modifier la modélisation de nos données afin que ces comportements soient visibles.

Préparation des séries temporelles

Comme mentionné dans la section 3.3.1, le choix des méthodes ou des techniques, pourraient impliquer des traitements supplémentaires des données. La figure 3.2 montre les étapes qui ont été effectuées pour construire les séries chronologiques.

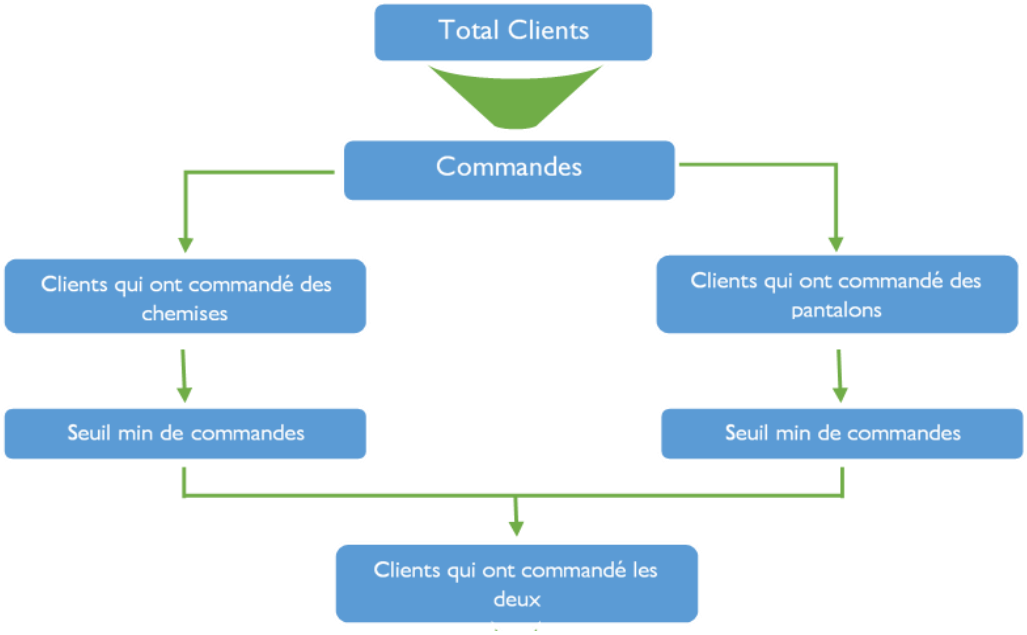


Figure 3.2 Préparation des données pour la construction des groupes de clients

Chaque ligne de la base de données originale, nous permet d'observer les détails d'un achat spécifique effectué par un client à un moment donné. Il y a des informations sur les dates auxquelles les transactions ont été effectuées, la quantité de produits qui ont été commandés et retournés, et les tailles des vêtements. Bien que toutes ces données soient pertinentes, il est nécessaire de transformer ces informations pour pouvoir construire des séries temporelles qui vont nous permettre d'observer l'évolution des mensurations des clients au fil du temps. Nous partirons du principe que la taille de vêtements commandés correspond aux données morphologiques de chaque individu.

Cependant, il faut souligner que les mesures ne sont pas les mêmes pour tous les vêtements, c'est-à-dire que les tailles des chemises et des pantalons ne se calculent pas de la même manière. Par conséquent, on devra construire un tableau de séries chronologiques spécifique à chaque mensuration.

Afin de construire des séries temporelles, le choix du niveau de granularité temporelle à laquelle nous désirons observer les données est important (El Ayeb et Agard, 2019). La granularité de la chronologie est la mesure du temps que l'on souhaite utiliser pour définir les périodes effectives des versions d'enregistrement. Par exemple, les périodes effectives peuvent être mesurées en années, mois ou jours (Del Mondo, 2011). Pour notre cas d'étude, on a choisi une granularité mensuelle. De plus, il faut trouver le moyen adéquat pour les agréger, parce qu'on est confronté à des séries temporelles intermittentes.

Segmentation des séries temporelles

Le but de cette étape est de diviser les séries chronologiques obtenues à l'étape précédente en groupes ayant des caractéristiques similaires. Tout d'abord, il est important de sélectionner la méthode la plus appropriée pour segmenter nos données. Dès lors, nous avons commencé à effectuer des tests avec les différents algorithmes dont dispose R, le logiciel utilisé, de cette façon nous procéderons au choix de l'algorithme le plus efficace pour les données dont nous disposons. Ainsi, l'algorithme TADpole est le plus approprié grâce à sa simplicité et il a l'avantage d'ignorer les points anormaux. Enfin, la segmentation est effectuée.

Au cours de cette étape, nous cherchons à détecter des tendances de variation similaires parmi les clusters obtenus à l'étape précédente, par exemple des modèles qui ont les mêmes tendances, afin de le regrouper dans des groupes plus importants. Une fois ces modèles identifiés, nous pouvons ainsi créer de nouveaux clusters avec ces données. Puis, nous analyserons la structure et les caractéristiques importantes des groupes créés. La composition de chacun des groupes est analysée, en tenant compte des informations clients spécifiques liées aux retours effectués. En plus de nous aider à observer les tendances, dans l'évolution des mensurations corporelles des individus, cette analyse nous permet également de comparer les groupes obtenus.

3.3.4 Construction des arbres de décision

Dans cette section, nous allons procéder à la construction de notre modèle de prédiction en utilisant les arbres de décision. Notre premier arbre est lié aux types de vêtements, où nous chercherons à prédire les caractéristiques des vêtements qui sont le plus souvent retournés. Le deuxième arbre est lié aux clients, dans ce cas nous chercherons à trouver les caractéristiques, et les tendances des clients qui effectuent des retours plus fréquemment. On va commencer par sélectionner les

variables pertinentes pour chaque cas. Ensuite, nous développerons nos arbres, pour cela nous effectuerons différents tests, en utilisant plusieurs variables et tailles d'échantillons.

3.3.5 Analyse et interprétation des résultats

L'analyse et l'interprétation des résultats, constituent l'étape qui permet la comparaison quantitative ou qualitative, des différentes solutions envisagées sur une base rationnelle. Elle doit également être explicitée, puisqu'il s'agit de la source même des conclusions, et des recommandations. Le principal objectif de cette étape est de comprendre les résultats obtenus à partir de notre modèle de prédiction. Cette analyse nous permettra d'appréhender les caractéristiques communes des clients qui font des retours. On sera aussi capable de connaître les types de produits qui sont retournés souvent, et leurs caractéristiques.

3.4 Conclusion

Dans ce chapitre, on a présenté la méthodologie suivie pour réaliser ce projet. Dans le chapitre suivant, on approfondira chacune des étapes mentionnées ci-dessus. Les résultats de notre étape de segmentation nous permettront de visualiser l'évolution des mensurations des clients, et de regrouper des tendances similaires. De plus, nous utiliserons les groupes obtenus pour réaliser le modèle de prédiction. Le modèle de prédiction développé pour la présente recherche permettra à notre partenaire de trouver les caractéristiques, et les tendances des clients qui effectuent la plupart des retours. On sera également en mesure de détecter les caractéristiques spécifiques des articles qui sont retournés.

CHAPITRE 4 CAS D'ÉTUDE

Dans ce chapitre, on commencera par présenter une brève introduction à la situation de notre partenaire industriel. Ensuite, on va développer chaque étape de la méthodologie appliquée dans le contexte pratique de l'étude. Finalement, on exposera les résultats obtenus, ainsi que leur analyse et les conclusions de cette étude.

4.1 Contexte

Fondée au Canada en 1993, notre partenaire industriel est une entreprise qui a pour mission de fournir la prestation de services pour la personnalisation, la conception et la production d'uniformes à différents clients organisationnels et gouvernementaux. La plupart de ses clients se trouvent au Canada, où plus de 325 000 personnes sont habillées par notre partenaire industriel. Ce dernier possède également des clients en Europe, en Australie, en Asie, en Afrique du Nord et au Moyen-Orient.

Les services fournis incluent aussi la recherche et le développement (R et D), le design, la production, l'approvisionnement, l'assurance-qualité et l'entreposage. L'entreprise soutient les industries locales du textile et de la confection vestimentaire en collaborant avec plusieurs fournisseurs à travers le Canada, et investit dans plusieurs projets de R et D tels que l'amélioration de la performance des vêtements sur le terrain, et le perfectionnement des méthodes de mesures dans les tests de laboratoire.

L'entreprise partenaire offre une plateforme en ligne pour les employés des entreprises soutenues, à partir de laquelle chaque individu peut commander des pièces de leurs uniformes, tout au long de l'année en utilisant un identifiant unique. Notre partenaire est également responsable de la livraison des pièces d'uniformes aux employés des entreprises clientes, en sous-traitant un service de transport.

Dans le présent projet de recherche, on limite notre étude à un groupe spécifique de clients choisi par notre partenaire. Chaque groupe de clients est associé à un sous-ensemble de produits disponibles différents, il n'est pas possible pour le client d'une entreprise A d'acheter une pièce d'uniforme de l'entreprise B, ainsi chaque contrat est indépendant, et peut être traité de manière indépendante. Le contrat analysé disposait d'un historique d'achat des données importantes, sur l'historique d'achats sur une période de 6 ans. Dans la prochaine étape, différentes analyses

statistiques seront présentées afin de mieux comprendre les informations fournies par l'entreprise partenaire.

4.2 Préparation de la base de données

Pour commencer à préparer la base de données, tout d'abord, il est nécessaire de l'importer dans un programme ou un logiciel doté des outils appropriés pour obtenir les résultats souhaités. Dans la présente étude, les analyses ont été effectuées avec le logiciel R.

La principale base de données qui a été utilisée contient un peu plus de 5 millions d'observations (plus de 100 000 clients au total), où figurent des informations sur le nombre des commandes et des retours de stock entre les années 2012 et 2018, ainsi que des informations supplémentaires telles que la date, le type de vêtement, la taille, le genre, etc.

On a commencé à faire le nettoyage à travers des filtres. Le premier filtrage a été réalisé pour ne sélectionner que les lignes contenant des retours, qui sont la base de cette recherche. On a aussi supprimé les lignes « Unisex » de la colonne Genre, car, selon les responsables de l'entreprise partenaire, le nombre de clients de ce groupe est minime, cette information a été vérifiée en faisant une analyse uniquement avec ce groupe de clients, où il a été constaté qu'ils représentent moins de 4% du total des clients. En outre, afin de mener ultérieurement une étude basée sur le sexe des clients, ces lignes ne contribueraient pas aux résultats. Finalement, il a également été nécessaire de modifier quelques colonnes, et de supprimer les lignes contenant des informations manquantes, ou des données considérées comme anormales. Les tableaux 4.1 et 4.2 présentent un résumé des variables, qui apportent des informations pertinentes pour les analyses, qui seront effectuées ultérieurement.

Tableau 4.1 Variables liées aux produits

Variable	Description
LOGISTIK NUMBER	Cette variable est simplement le type de produit : une chemise, un pantalon, des chaussettes, etc.
PROD_GROUP	Cette variable contient les spécifications des produits comme la couleur, les poches, les manches, etc.
GMT_SEX	Cette variable indique si le produit est pour les femmes, les hommes, ou s'il est pour les deux.
GMT_SEASON	Cette variable indique si le produit est pour la saison d'hiver, d'été ou les deux (hiver- été).
ENTCODE	Cette variable spécifie l'entité pour laquelle le produit a été fait.
SEASON	Cette variable indique l'année de fabrication des vêtements.

Tableau 4.2 Variables liées aux clients

Variable	Description
CLIENT_SEX	Cette variable indique le sexe du client.
ENTCODE_DESCRIPTION	Cette variable indique l'entité à laquelle le client appartient.
RANK	Cette variable indique le rang du client dans les fonctions qu'il exerce.
GROUP	Cette variable indique le groupe auquel appartient le client.
POSTAL_CODE	Cette variable indique l'emplacement du client.

4.3 Analyses statistiques

L'analyse statistique des données permet de mieux visualiser les informations disponibles, pour nous aider à prendre des décisions.

4.3.1 Analyse dans le temps

Une fois nos modifications effectuées, il a été possible de porter ces résultats sur un premier tableau, qui nous permet de mieux connaître la répartition des clients, et leur pourcentage de retours et de commandes.

Tableau 4.3 Distribution de la base de données

	Hommes	**Femmes**	
% Personnes	82, 41%	17,59%	100%
% Commandes	82,97%	17,03%	100%
% Retours	71,12%	28,88%	100%
% Retours/Commandes	1,40%	2,77%	

Comme le montre le tableau 4.3, un peu plus de 80% des clients de notre base de données sont des hommes, et presque 20% des femmes. Logiquement, la plupart des commandes et des retours sont effectués par des hommes, en raison de leur présence majoritaire dans la base. Cependant, lorsque nous avons calculé la proportion de retours par rapport aux commandes faites par chaque sexe, il ressort que proportionnellement les femmes font plus de retours (2,77%) que les hommes (1,40%).

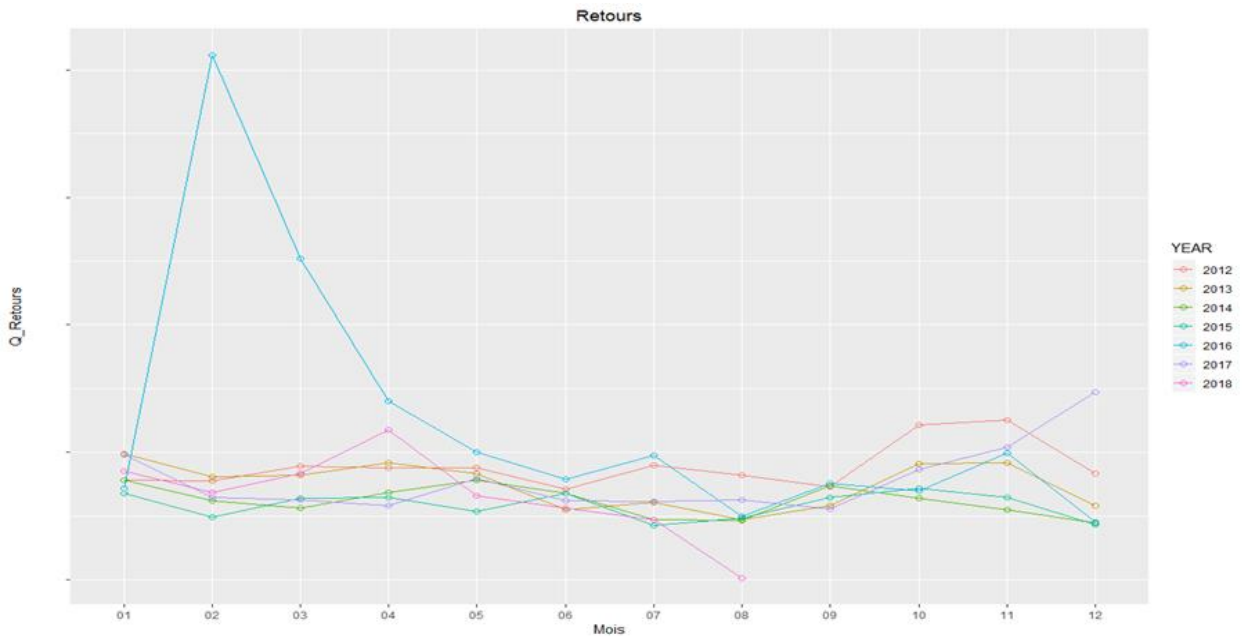


Figure 4.1 Retours effectués pendant la période d'étude.

Poursuivant les analyses, on a décidé de se concentrer sur les retours. La figure 4.1 nous montre les quantités de retours faits par mois pendant toute notre période d'étude. Nous observons deux points importants. Le tableau 4.4 résume les résultats de la figure 4.1, on observe que les quantités de retours diminuent d'année en année, sauf pour 2016, qui représente une situation particulière où le nombre de retours augmente énormément.

Tableau 4.4 Pourcentage de retours par année

	% Retours
2012	16,75%
2013	14,12%
2014	11,64%
2015	11,04%
2016	23,42%
2017	14,79%
2018	8,24%

C'est principalement en raison des retours du mois de février 2016, que le volume de vêtements retournés a été excessif par rapport à toute autre période au cours des 6 années étudiées. Dans ce sens, on a considéré pertinent de procéder à une analyse plus approfondie de ce point, afin de connaître et comprendre les événements, qui auraient pu provoquer ce résultat.

On a commencé par examiner les groupes de vêtements et la raison pour laquelle ils ont été retournés. Pour cela la figure 4.2 montre les raisons des retours avec leur montant respectif. De plus, la légende nous permet d'identifier les types de vêtements retournés.

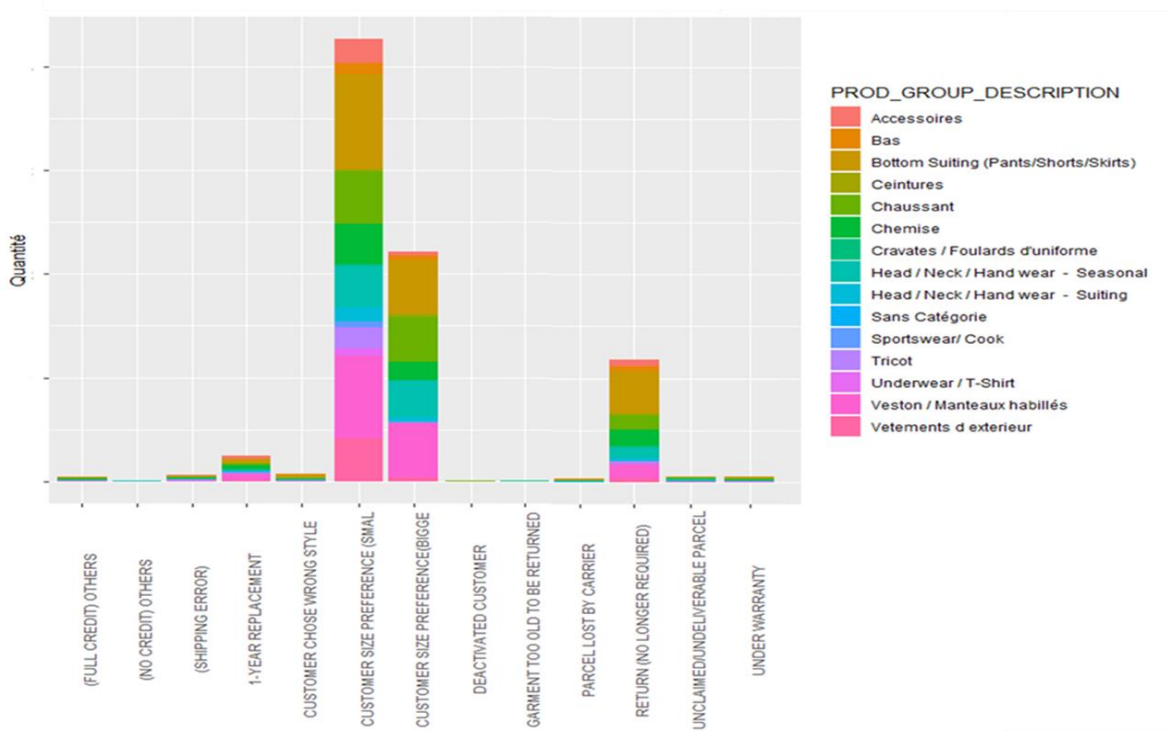


Figure 4.2 Raisons de retour par vêtement en 2016

L'étude de la figure 4.2 montre que les vêtements avec le plus grand nombre de retours en février 2016, étaient les badges et les items métalliques, les vêtements pour le bas du corps et les vestes. Nous pouvons également distinguer que les raisons de retour les plus courantes sont liées à la taille (parfois plus petite ou plus grande).

4.3.2 Analyse par sexe

Ensuite, la base de données a été divisée par sexe pour voir s'il existait des variations dans le comportement des clients. Dans les deux cas, le pic de février 2016 est présent, on a donc approfondi l'analyse. Ce qui donne les graphiques 4.3 et 4.4, où l'on constate que les vêtements les plus retournés par les femmes sont les jupes et les badges. En revanche, dans le cas des hommes, bien que les pantalons et les badges soient également retournés en grand nombre, on retrouve d'autres vêtements tels que des accessoires, des manteaux et des vêtements pour les mains et le cou.

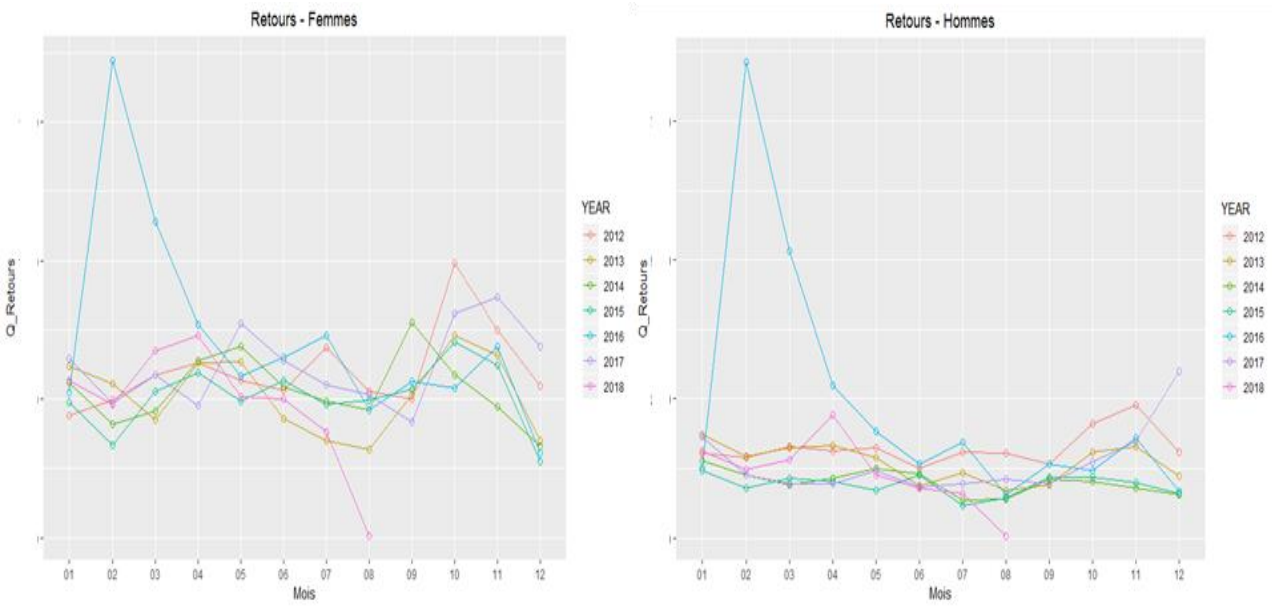


Figure 4.3 Nombre de produits retournés pendant la période d'étude divisée par sexe

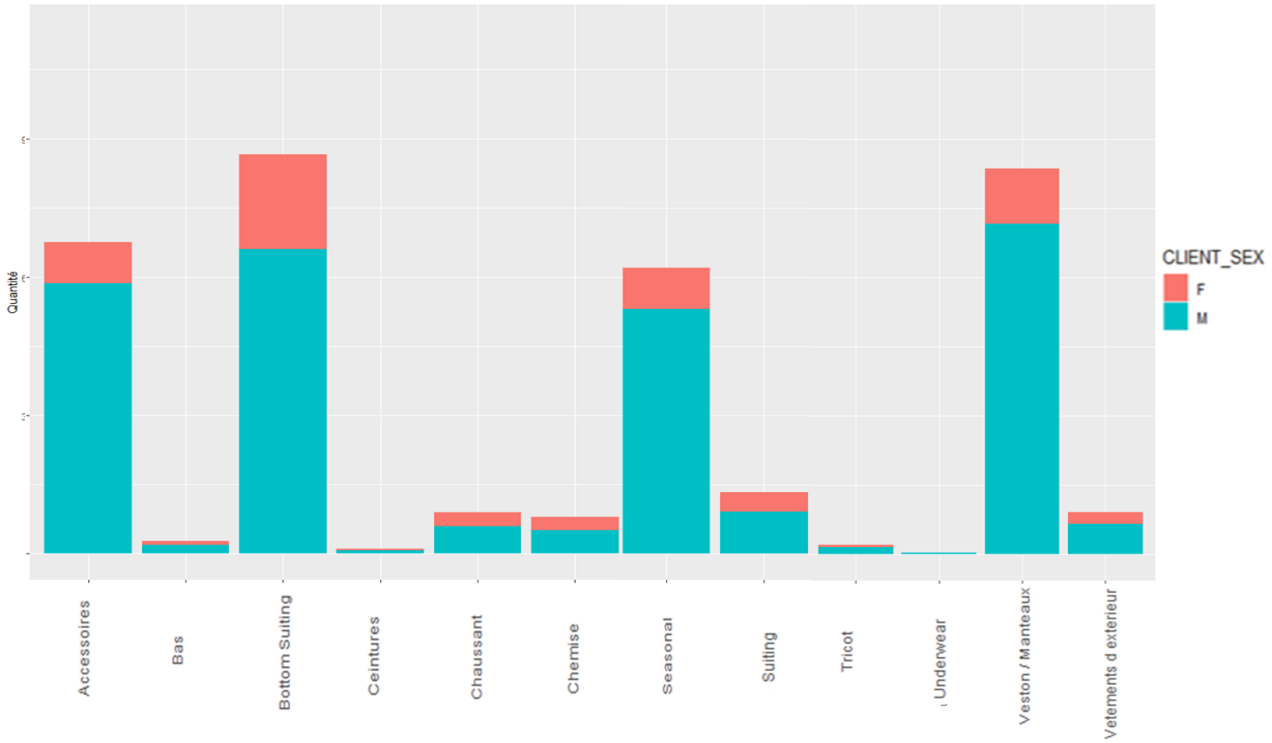


Figure 4.4 Retours effectués en 2016 divisés par sexe

De la même manière, une analyse a été effectuée pour chacun des trois groupes ou entités, et ses subdivisions afin d'avoir plus en détail les informations. Contrairement aux analyses précédentes, les comportements de ces groupes étaient plus aléatoires. Les résultats obtenus étaient assez variés.

Parmi les cas les plus notables, figure l'entité 1A, l'un des cas peut-être les plus aléatoires, dont le plus grand volume de retours était en novembre 2012 (figure 4.5). Comme nous pouvons l'apprécier (figure 4.5), il y a plusieurs pics qui représentent des grandes quantités de retours qui ont été faits aux différentes dates, et on ne voit pas le pic de Février en 2016. Par conséquent, nous avons décidé de prendre cette date et d l'examiner. Nous pouvons observer les résultats sur la figure 4.6. Le plus grand nombre de vêtements pour le bas du corps, et toutes les chaussures ont été retournés par les femmes. Ce résultat est intéressant, car la plupart des clients sont des hommes.

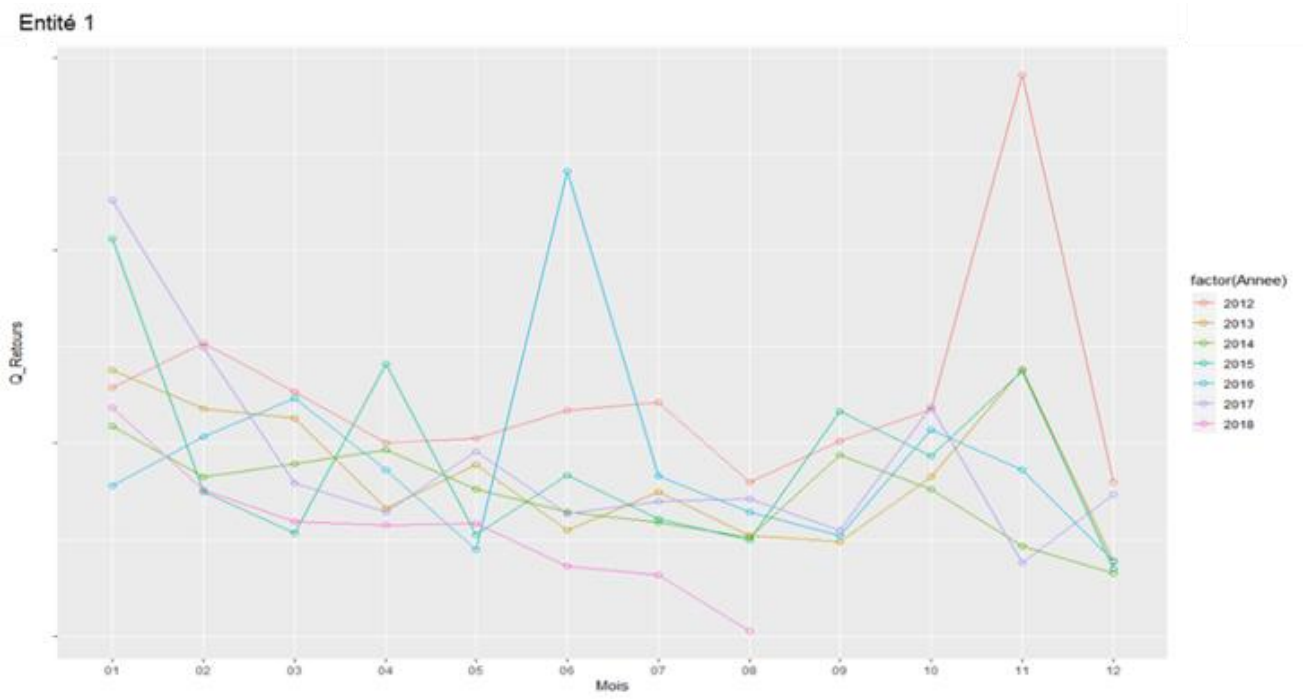


Figure 4.5 Retours faits par l'entité 1A

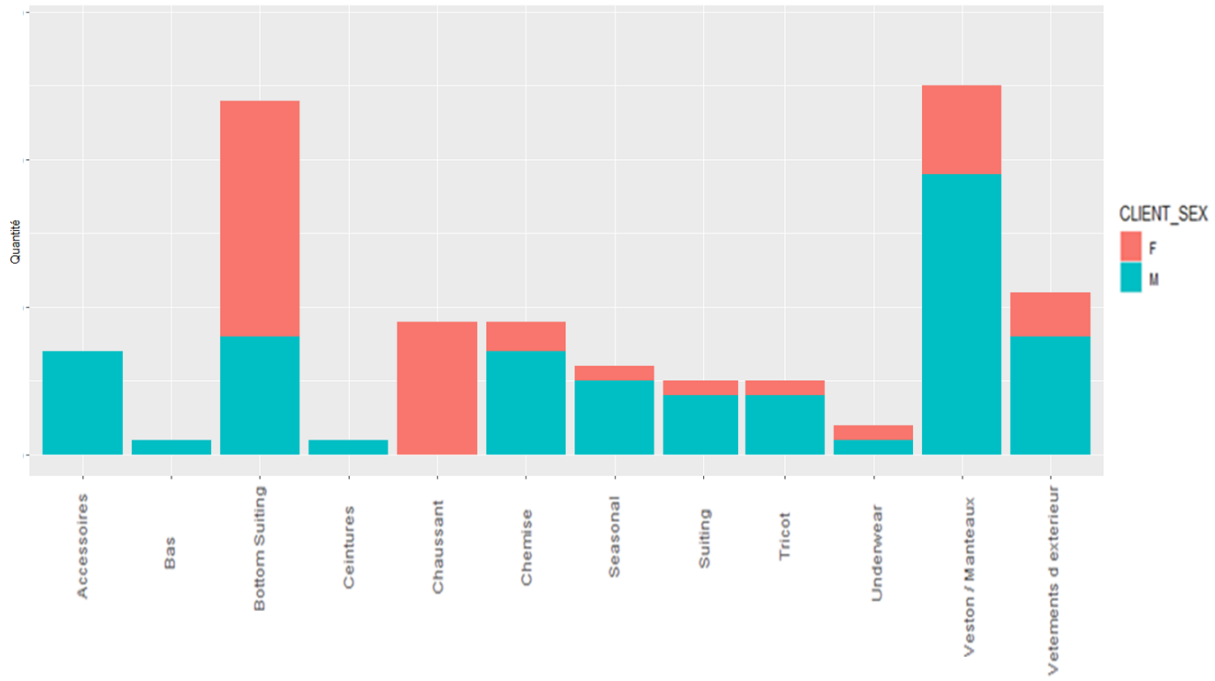


Figure 4.6 Retours effectués par les membres de l'entité 1A

L'une des observations les plus importantes faites lors de l'obtention de nos résultats a été l'énorme quantité de retours de badges, et d'items métalliques pour des raisons de mauvaise taille. Cependant, en consultation avec notre partenaire, les gestionnaires ont allégué que ce phénomène est normal, car les vestes ont un certain nombre de badges, donc lorsqu'un client commande ou retourne une veste, le nombre de badges inclus sera également saisi avec la même raison de retour.

De plus, on a pu constater avec l'entreprise que les volumes de retours massifs en 2016, étaient dus à l'introduction d'un nouveau type de veste en 2015. Cela expliquerait également la grande quantité d'items métalliques retournés en 2016, liés à ces vestes. Par la suite, les retours des badges ne seront pas considérés, ils n'apportent pas d'information sur les tailles des vêtements, leur retour est simplement lié à leur association avec les vêtements commandés.

4.3.3 Analyse par type de vêtement

On a décidé de réaliser une dernière analyse globale en prenant tous les vêtements, sauf les badges, pour connaître ceux qui ont le plus de retours. Les résultats s'observent sur la figure 4.7 et 4.8, où clairement on trouve que les chemises, les pantalons-jupes et les vestes sont les produits avec la plus grande quantité de retours.

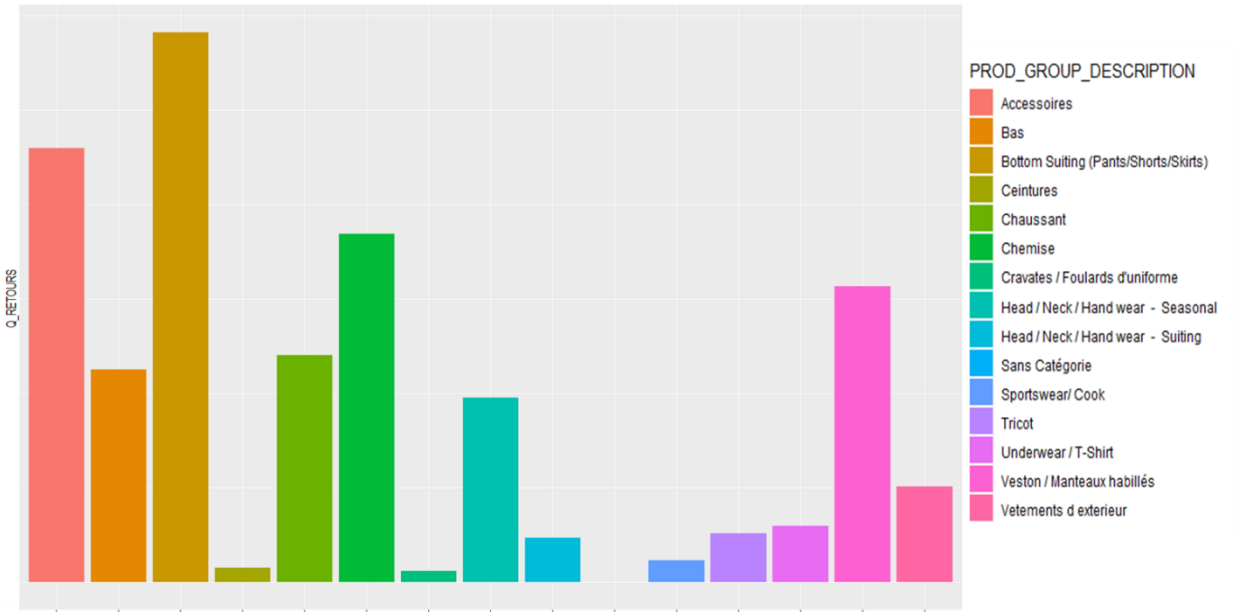


Figure 4.7 Quantité de retours - type de produit

La figure 4.8 nous permet d'observer les vêtements et la raison de leur retour. Le résultat montre clairement que la raison principale du retour des vêtements est la taille.

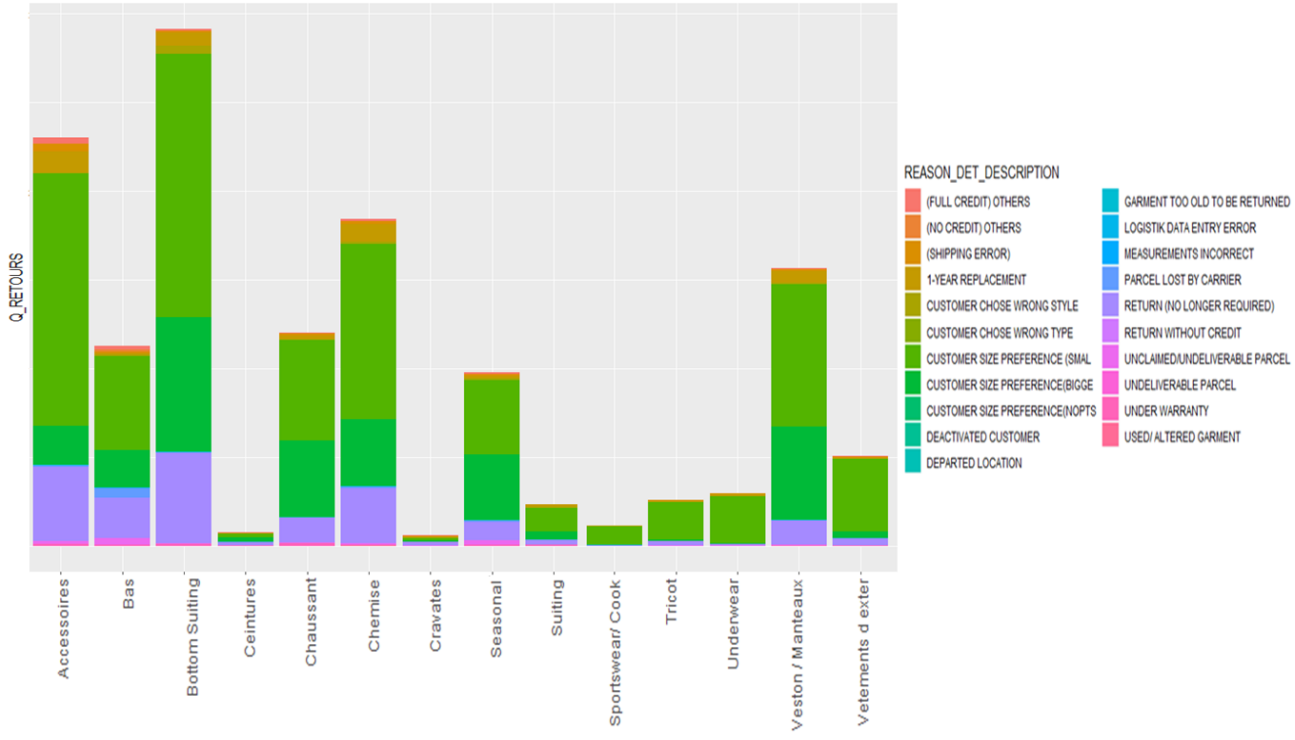


Figure 4.8 Raisons de retours par vêtement

4.4 Construction des groupes de clients

L'objectif de cette étape est de regrouper les clients en fonction de l'évolution de leurs mesures corporelles. Par conséquent, on a décidé d'effectuer une segmentation de l'évolution des tailles de vêtements commandés, qui nous permettra d'obtenir des comportements similaires parmi les individus impliqués.

4.4.1 Préparation des données

Afin d'effectuer la segmentation, on a procédé à prendre la base préalablement filtrée, où les lignes des clients « Unisex » ont été enlevées, et où seules les commandes (avec retours) ont été prises en compte. Ensuite, nous avons procédé à la division de la base de données en deux. La première base contient les commandes de chemises d'hiver ; tandis que la deuxième base contient les commandes de pantalons de la même saison. Ces deux vêtements ont été sélectionnés en raison du grand nombre de commandes et de retours qu'ils présentent.

Après avoir effectué une analyse sur les deux bases, on a constaté que certains clients n'ont fait qu'une seule commande pendant toute la période d'étude. Pour cette raison, il a été nécessaire de fixer un seuil minimum de 4 commandes par client. De cette façon, il sera possible de visualiser et d'estimer, une possible variation de taille de l'individu sur la période.

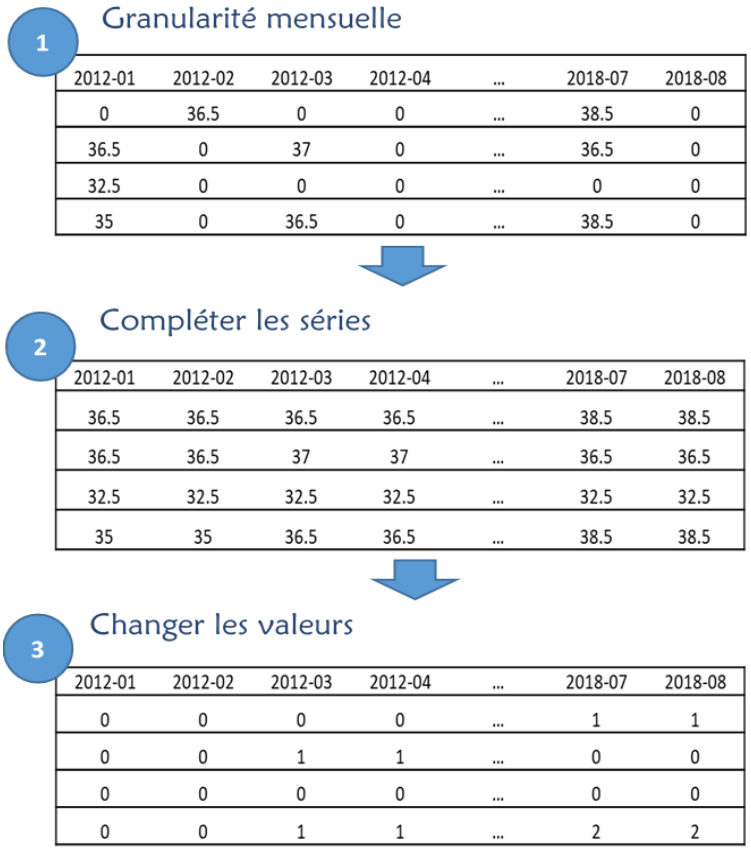
Pour la fabrication de chaque vêtement, différentes mesures corporelles sont prises en compte, cependant les plus courantes sont : le tour du cou, de la poitrine, de la taille et des hanches. En ce sens, on a décidé de travailler avec chacune de ces quatre mesures, puisque les vêtements sélectionnés pour l'analyse contiennent ces informations. Enfin, on a procédé à faire une jointure, entre le tableau de chemises et celui de pantalons, afin d'obtenir les clients auprès desquels les quatre principales mesures sont connues.

4.4.2 Construction des séries temporelles

Bien que la base de données obtenue à l'étape précédente, nous fournisse des informations pertinentes telles que les mesures principales, et les dates des commandes, la construction de séries temporelles des tailles de vêtement commandées est essentielle, pour pouvoir rendre visible l'évolution de ces mesures dans le temps. Cependant, une fois créés, nous sommes confrontés à un problème dû à la granularité quotidienne qu'ils présentent. Les commandes ont été saisies avec une

date précise, ce qui nous donne des données très intermittentes. Pour surmonter ce problème, on a décidé de suivre une méthode qui a été utilisée auparavant (El Ayeb et Agard, 2019), où les séries temporelles ont été agrégées au niveau mensuel. De plus, puisque le changement de la granularité n'améliore pas complètement le problème d'intermittence, on va aussi prendre la commande initiale pour chaque individu, et chaque fois que cette personne ne passe pas une nouvelle commande, on va considérer que sa taille reste constante.

Afin d'obtenir de meilleurs résultats, nous avons également décidé de standardiser les données, donc on suppose que chaque client commence la période avec une taille égale à zéro. Si la prochaine commande de cette personne est plus grande, on augmentera d'une unité; et si, au contraire, le client commande une taille plus petite, on soustraira une unité. On peut observer cette procédure dans la figure 4.9. L'intérêt n'est pas de prédire la taille d'un individu, mais de voir si cette taille évolue, et si elle entraine des retours de produits plus fréquents.



2012-01	2012-02	2012-03	2012-04	...	2018-07	2018-08
0	36.5	0	0	...	38.5	0
36.5	0	37	0	...	36.5	0
32.5	0	0	0	...	0	0
35	0	36.5	0	...	38.5	0

2012-01	2012-02	2012-03	2012-04	...	2018-07	2018-08
36.5	36.5	36.5	36.5	...	38.5	38.5
36.5	36.5	37	37	...	36.5	36.5
32.5	32.5	32.5	32.5	...	32.5	32.5
35	35	36.5	36.5	...	38.5	38.5

2012-01	2012-02	2012-03	2012-04	...	2018-07	2018-08
0	0	0	0	...	1	1
0	0	1	1	...	0	0
0	0	0	0	...	0	0
0	0	1	1	...	2	2

Figure 4.9 Étapes suivies pour construire les séries chronologiques

De cette manière, on a été capable d'obtenir des séries chronologiques continues pour les quatre mesures principales, voici par exemple, ce qu'on obtient pour un client particulier.

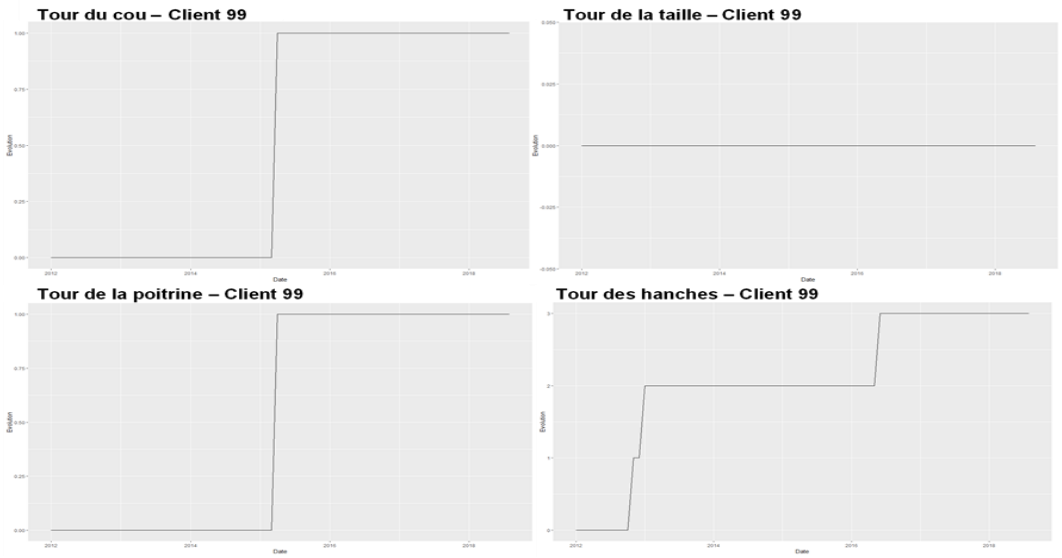


Figure 4.10 Évolution des quatre mesures principales du client 99.

Comme on l'observe sur la figure 4.10, le comportement de l'évolution des tailles d'un client peut suivre un même patron. Dans le cas du client 99, les mesures « tour de cou » et « tour de poitrine » présentent la même variation. Cependant, si nous observons les mesures « tour de taille » et « tour des hanches », on se rend compte que les comportements sont complètement différents. Le client 99 est considéré comme stable dans la mesure « tour de taille », mais pas dans la mesure « tour des hanches ».

4.4.3 Segmentation des clients

Cette étape, consiste à diviser les clients en groupes ayant des comportements similaires, en fonction de l'évolution de leurs mensurations. Avant de procéder à la segmentation, on a aussi décidé de mettre à côté les clients qui ont une courbe stable, c'est-à-dire ceux qui ont commandé la même taille tout au long de la période d'étude, afin de se concentrer sur ceux qui présentent des changements. On a ensuite procédé au choix de la méthode de segmentation, et on utilise la distance Dynamic Time Warping (DTW). Cette technique est très populaire pour comparer les séries temporelles, car non seulement la valeur de dissimilarité entre deux séries est évaluée, mais également l'alignement optimal entre elles est déterminé, en les faisant correspondre de manière non linéaire au moyen de contractions de séries, et de dilatation dans l'axe du temps (Zhao et Itti,

2018). Par conséquent, cet appariement permet de trouver des régions équivalentes entre les séries, et de trouver leur similitude.

4.4.3.1 Choix de l'algorithme de segmentation

L'algorithme TADPole a été choisi pour faire la segmentation. Il adopte un cadre de segmentation nouveau, inspirée de l'algorithme des pics de densité (DP) développée par Rodriguez et Laio (2014), et l'adapte aux segmentations de séries temporelles avec DTW. L'algorithme cherche les séries qui se trouvent dans des zones denses, c'est-à-dire, les séries qui ont beaucoup de voisins, et ils sont considérées comme des centres de gravité. Il est nécessaire de définir la distance de coupure (cc), toute série ayant une distance inférieure à cette distance, est considérée comme voisine (Begum et al, 2015).

Il y a deux principales raisons pour lesquelles on a choisi cet algorithme, sa capacité d'ignorer les points des données anormaux et sa simplicité. Il est nécessaire de définir seulement deux paramètres, la distance de coupure et le nombre de clusters (k). Il convient de noter que ce deuxième paramètre n'est pas nécessaire, car l'algorithme utilise une heuristique de recherche de coude et l'indice de silhouette pour suggérer un nombre optimal de groupes k. Il est aussi nécessaire de mettre une contrainte de fenêtrage. La méthode TADPole utilise une fenêtre centrée, c'est-à-dire que les calculs s'attendent à une valeur qui représente la distance entre le point considéré, et l'un des bords de la fenêtre.

De par nature, la méthode de DTW est basée sur la notion de distance, et ne fait pas la différence si cette distance est positive ou négative, seule la valeur absolue de la distance est retenue. Il apparaît alors des résultats tels que ceux illustrés à la figure 4.11, où un écart positif (personne qui grossit dans notre cas) est regroupée avec un écart négatif (personne qui maigri). Cela ne fait pas de sens dans notre cas d'étude, et différents essais ont été réalisés pour le choix de la contrainte de fenêtrage.

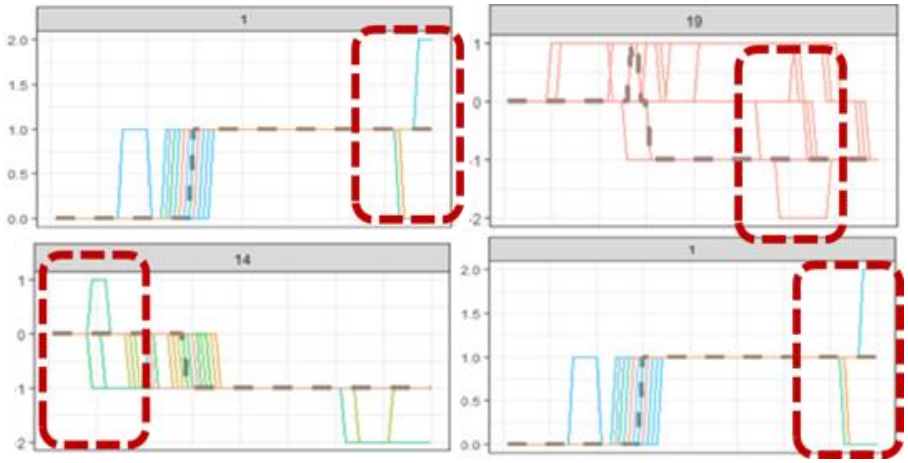


Figure 4.11 Exemples – problèmes de segmentation

Comme le montre la figure précédente (figure 4.11), les variations ascendantes et descendantes, peuvent être classifiées comme des comportements similaires, et même considérées comme une évolution linéaire ou stable.

Comme mentionné ci-dessus, l'objectif de notre segmentation est de trouver des tendances similaires chez nos clients, pour cette raison, afin d'éviter le problème mentionné précédemment, on a effectué plusieurs tests pour chacun de nos paramètres, jusqu'à l'obtention de résultats optimaux. Il convient de mentionner qu'on a commencé par prendre le nombre de clusters (K) suggéré par l'algorithme (cette valeur étant différente pour chaque mesure). Les valeurs utilisées dans notre cas d'étude se trouvent dans le tableau suivant (tableau 4.5).

Tableau 4.5 Paramètres utilisés pour la segmentation

	Distance de coupure (cc)	**K**	**Contrainte de fenêtrage**
Mesure 1	15	25	40
Mesure 2	6	25	40
Mesure 3	15	30	50
Mesure 4	20	25	45

La distance de coupure est la distance d'un point au point le plus proche ayant une densité élevée, et K représente le nombre de clusters. De cette façon, on a été capable d'obtenir les résultats souhaités pour chacune des mesures. La figure 4.12 montre la segmentation effectuée pour la mesure « tour de cou ». La figure montre 25 clusters, dans chaque cluster, la ligne pointillée représente la variation de la mesure Tour de cou « moyenne »

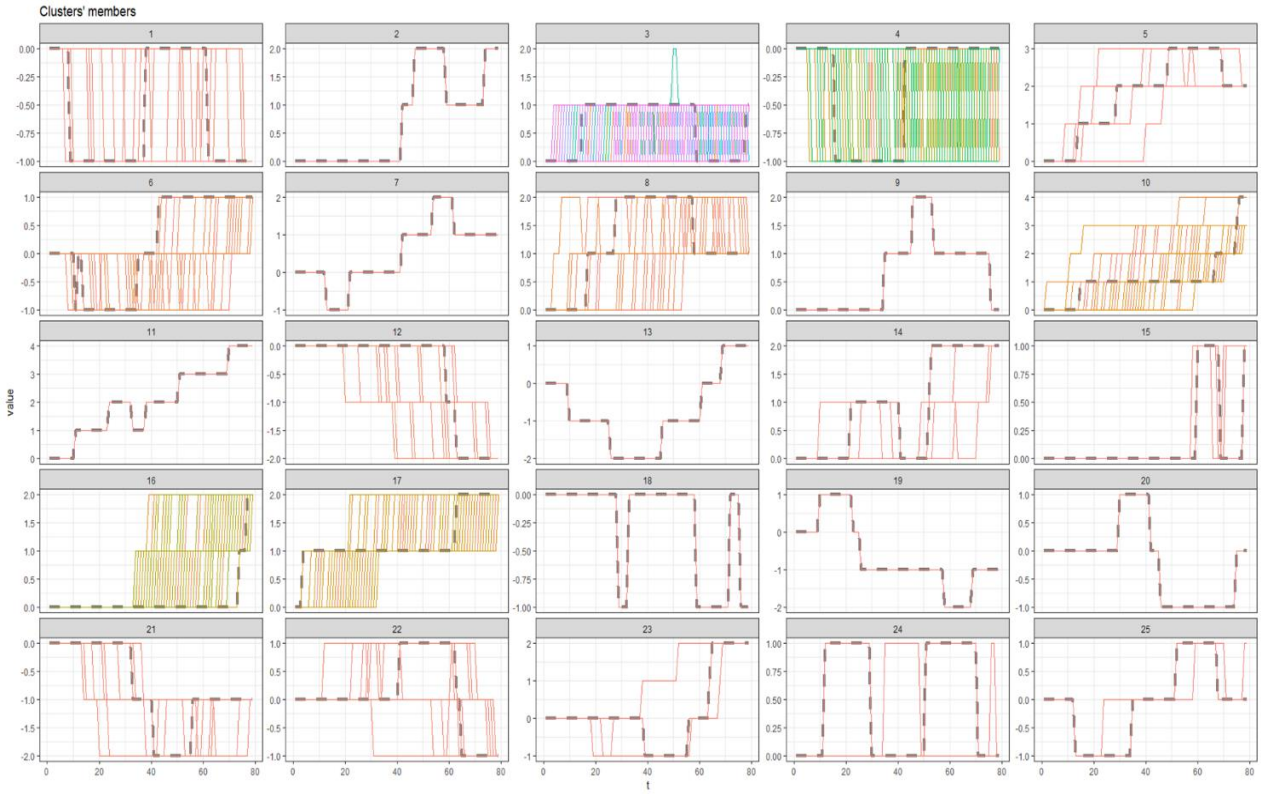


Figure 4.12 Résultats de la segmentation pour la mesure 1 : tour de cou

On observe que différents clusters ont des comportements similaires, par exemples les clusters 1,7, 18 et 25 ont le comportement suivant : leurs courbes commencent par descendre, puis elles varient et à la fin de la période, elles finissent par descendre.

4.4.4 Regroupement des clusters

Étant donné que les différents groupes obtenus, pour chaque segmentation montrent des comportements similaires, à cette étape il a été décidé de regrouper manuellement les clusters, qui présentent la même tendance dans un groupe plus général, avec un nouveau numéro.

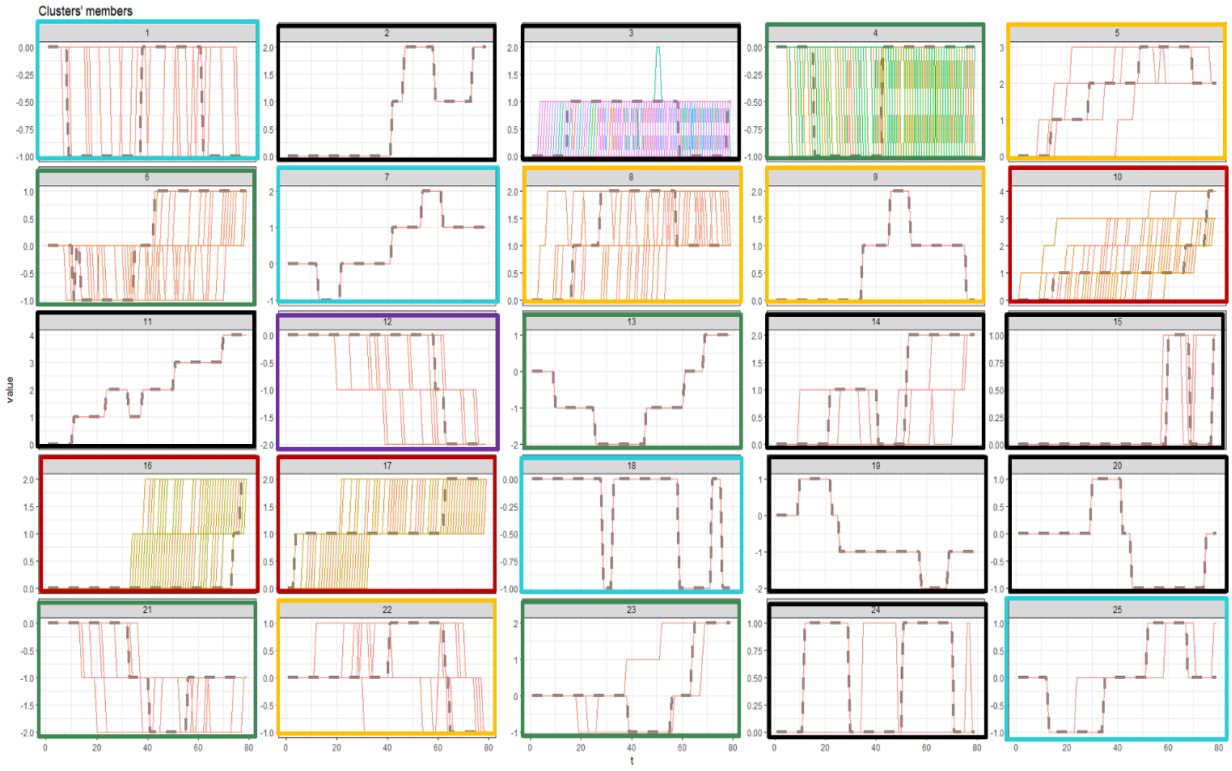


Figure 4.13 Regroupement de clusters – Mesure 1 : tour du cou

Dans la figure 4.13, on a encadré avec la même couleur les clusters qui seront réunis dans un même groupe plus général. On est donc capable d'observer les membres du premier « nouveau cluster 1 » en bleu clair, qui inclut les grappes 1, 7, 18 et 25. Ce premier groupe présente des personnes avec un comportement variable, les gens commencent par commander des tailles plus petites, puis ils montent et descendent et à la fin de la période d'études, ils commandent des tailles plus petites que leur taille précédente. Le « nouveau cluster 2 » en noir, contient les gens qui commencent par commander des tailles plus grandes que celle qu'ils avaient au début, puis la tendance de la courbe monte et descend, et à la fin la courbe monte. En vert, on apprécie le « nouveau cluster 3 », où on trouve les clients qui montrent une tendance qui commence par descendre, mais qui monte à la fin. Le « nouveau cluster 4 » en jaune, contrairement au nouveau cluster 3, contient les clients avec une tendance qui commence par monter, mais qui descend après. On voit les membres du « nouveau cluster 5 » en rouge, dont leur tendance ne fait que monter. Finalement, en violet, le « nouveau cluster 6 » contient les personnes qui ont tendance à commander des tailles de plus en plus petites. On passe ainsi de 25 à 6 groupes.

La figure 4.14 nous montre les regroupements faits pour le « tour de poitrine »

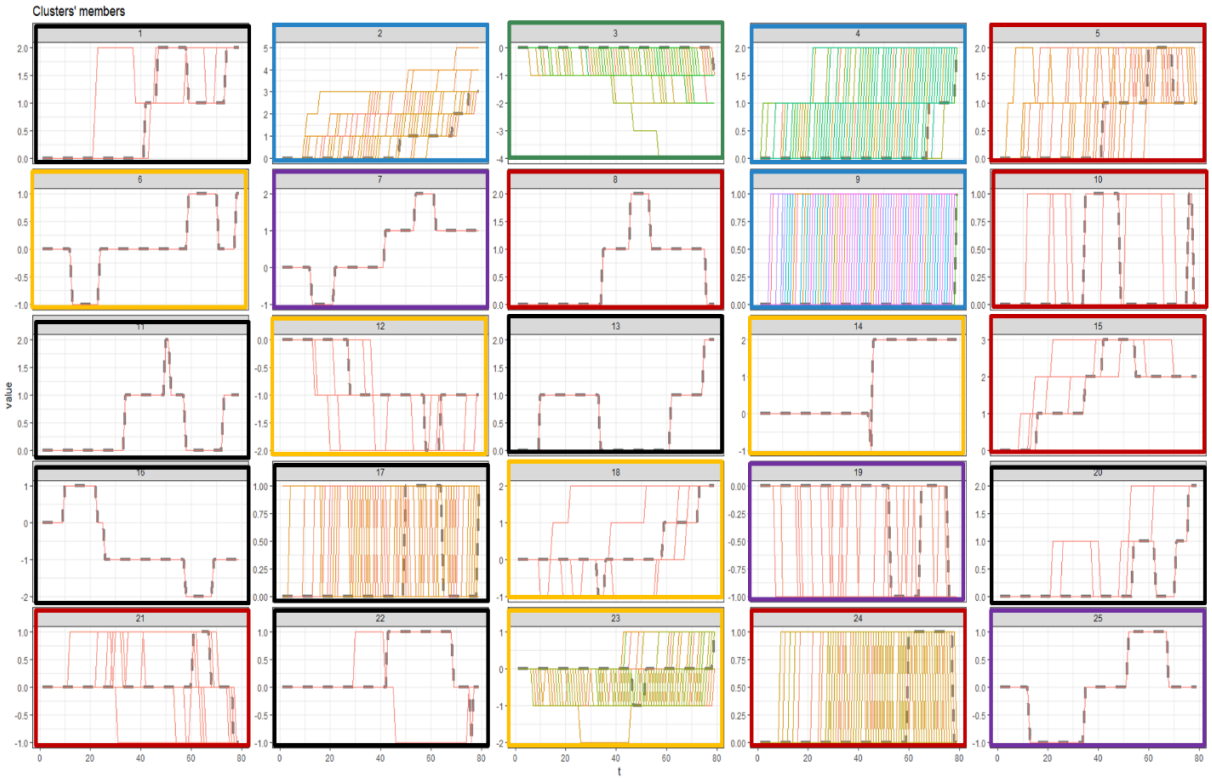


Figure 4.14 Regroupement de clusters – Mesure 2 : tour de la poitrine

Le « nouveau cluster 1 » en couleur noir inclut les groupes 1, 11, 16, 17 et 22. Ce premier nouveau groupe présente des personnes qui commencent par commander des tailles plus grandes, puis ils commandent une taille plus petite, mais ils finissent par commander une taille plus grande. On voit le « nouveau cluster 2 » en bleu, qui contient les gens qui ne font que commander des tailles de plus en plus grandes. Contrairement au nouveau cluster 2, le troisième nouveau cluster comprend des personnes qui ont tendance à commander des tailles de plus en plus petites. Le « nouveau cluster 4 » en rouge, contient les clients avec un comportement similaire au premier nouveau cluster, puisque la courbe de comportement de ces individus commence par monter, mais à la fin elle descend. Les membres du « nouveau cluster 5 » sont entourés en couleur jaune, ils possèdent une tendance qui commence par descendre et qui monte à la fin. Pour finir, on a en couleur violet, le « nouveau cluster 6 » qui présente des clients qui commencent par commander des tailles plus petites, puis qui commandent une taille plus grande, mais qui finalement finissent par commander une taille plus petite que la précédente. On passe aussi de 25 à 6 groupes.

La segmentation du tour de taille donne cette fois 30 groupes (Figure 4.15)

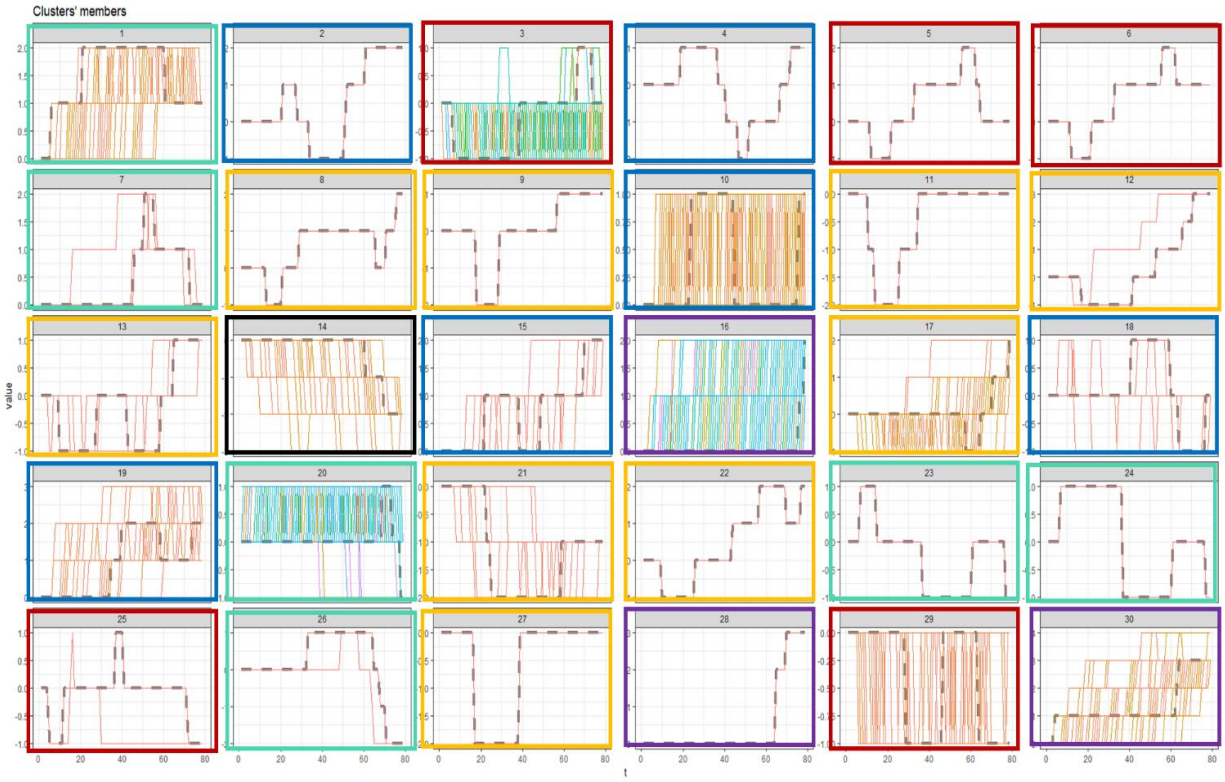


Figure 4.15 Regroupement de clusters – Mesure 3 : tour de la taille

Le « premier nouveau cluster » en cyan nous montre les clients qui commencent par commander des tailles plus grandes que leur taille initiale, mais qui après ne font que commander des tailles plus petites. Ensuite, on a le « nouveau cluster 2 » en bleu, qui contient les personnes qui commencent par commander des tailles plus grandes, puis commandent des tailles plus petites, mais finissent par commander des tailles plus grandes. On voit les membres du « nouveau cluster 3 » en rouge, la tendance de ce cluster commence par descendre, puis la courbe monte et à la fin elle descend. Le « nouveau cluster 4 » en jaune contient les clients avec une tendance similaire aux gens du nouveau cluster 3, puisque leur courbe commence par descendre, mais après finit par monter. Les membres du « nouveau cluster 5 » en couleur noire sont les personnes que ne font que commander des tailles plus petites chaque fois. Finalement, en violet, le « nouveau cluster 6 » contient les personnes qui ont tendance à commander des tailles de plus en plus grandes.

Pour terminer, la figure 4.16 montre les résultats obtenus pour la quatrième et dernière mesure, le tour des hanches.

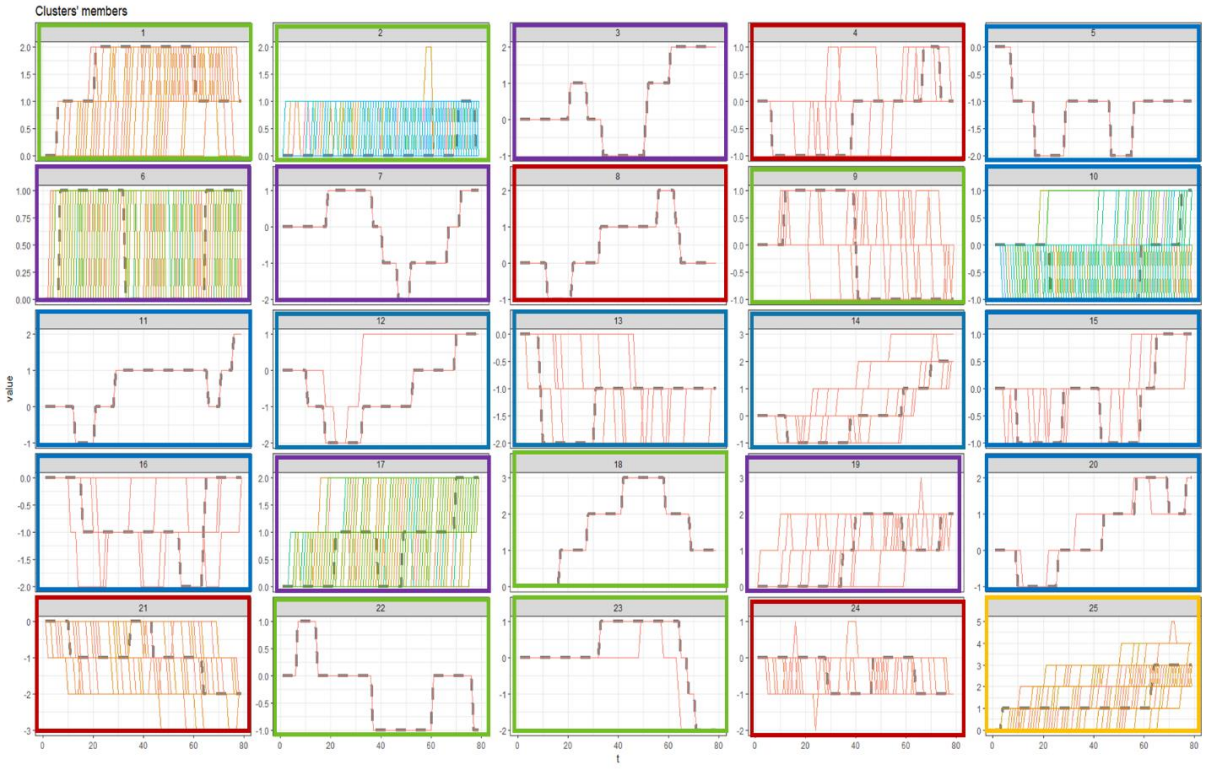


Figure 4.16 Regroupement de clusters – Mesure 4 : Tour des hanches

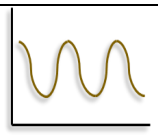
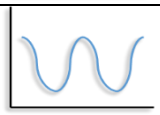
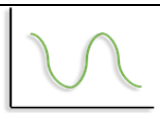
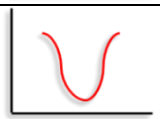
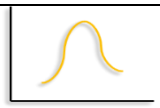
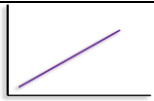
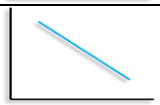
On est donc capable d'observer les membres du premier « nouveau cluster 1 » en vert qui inclut les personnes qui commencent par commander des tailles plus grandes, mais après ils commandent des tailles plus petites. Le « nouveau cluster 2 » en violet, contient les gens qui ont une tendance qui commence par monter, puis descend une ou deux fois, mais finit par monter. En rouge, on peut apprécier le « nouveau cluster 3 », où l'on trouve les clients qui commencent par commander des tailles inférieures à leur taille initiale, puis ils commandent des tailles plus grandes, mais à la fin commandent des tailles plus petites. Les membres du « nouveau cluster 4 » peuvent être reconnus par la couleur noire. Ce groupe contient les clients avec une tendance qui descend et qui monte après. On voit les membres du « nouveau cluster 5 » en bleu, dont leur tendance apparait variable, avec une courbe qui commence par monter, puis descendre, et finit par monter. Finalement, en jaune, le « nouveau cluster 6 » contient les personnes qui ne font que commander des tailles plus grandes.

4.4.5 Analyse des groupes

Après avoir effectué les tests nécessaires et obtenu les groupes générés dans la section précédente, des analyses permettent de trouver les caractéristiques des clients qui composent chacun des clusters. Il faut noter que pour cette phase, on a également ajouté un dernier cluster pour chaque mesure, qui contient les clients stables, qui avaient été retirés de l'analyse avant la construction des clusters.

Dans le tableau 4.6, la colonne « Tendance » exprime l'évolution de la morphologie moyenne du groupe de clients. Différentes évolutions ont été observées, et sont représentées par les légendes suivantes (Tableau 4.6)

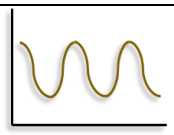
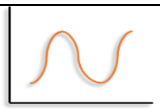
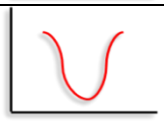
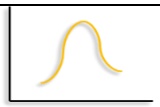
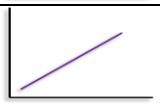
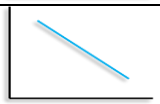
Tableau 4.6 Tendance des évolutions observées

	Tendance	**Évolution observée**
1		Comportement variable, une évolution qui commence par descendre, puis la courbe oscille et fini par diminuer.
2		Comportement variable qui commence par descendre, puis la courbe oscille et fini par monter.
3		Comportement variable qui commence par monter, puis a tendance à diminuer et finalement augmenter à nouveau.
4		Comportement variable, une évolution qui commence par descendre.
5		Comportement qui commence par descendre mais qui remonte à la fin.
6		Comportement qui commence par monter mais qui finit par diminuer à nouveau.
7		Comportement qui ne fait qu'augmenter. Les mesures deviennent plus grandes.
8		Comportement qui ne fait que diminuer. Les mesures deviennent plus petites.
9		Comportement stable. Les mesures n'ont jamais changé.

Le tableau 4.6 fait ressortir 9 comportements observés, chaque légende représente un type d'évolution particulier. Par exemple, la première ligne du tableau montre un comportement, où la taille des individus oscille et fini par diminuer.

Dans chacun des tableaux ci-dessous, sont représentés, le numéro du cluster, un croquis sur la tendance observée, le pourcentage de clients concernés, la proportion de femmes et homme répartis entre les différents groupes.

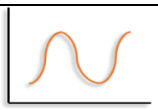
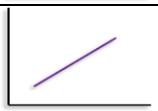
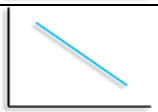
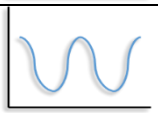
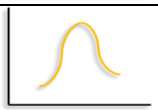
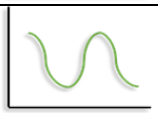
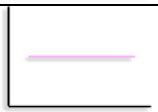
Tableau 4.7 Analyse de résultats – Tour de cou

	Cluster	Tendance	%Total	% Femmes	% Hommes	% Retours
1	Clus1		0,91%	0,64%	1,00%	1,31%
2	Clus2		32,88%	32,48%	33,00%	39,77%
3	Clus3		15,73%	17,62%	15,14%	19,32%
4	Clus4		2,68%	1,70%	2,99%	0,38%
5	Clus5		15,28%	15,07%	15,34%	8,63%
6	Clus6		0,86%	1,49%	0,66%	2,06%
7	Clus7		31,66%	31,00%	31,87%	28,52%
			100,00%	100,00%	100,00%	100,00%

Le tableau 4.7 présente les résultats obtenus pour le tour du cou. Comme on peut le voir, les groupes 2 et 7 contiennent la plupart des clients (32,88% et 31,66% respectivement), c'est donc le groupe avec le plus retours. Il est important de noter qu'au sein du cluster 2, on trouve des clients qui commencent par commander des tailles plus grandes que celles qu'ils avaient au début, puis la

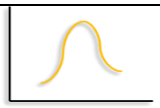
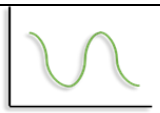
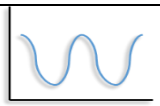
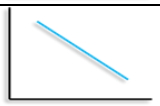
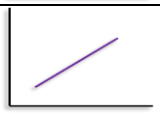
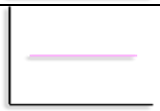
tendance de leur courbe monte et descend, et à la fin la courbe monte, tandis que le cluster 7 contient les gens stables. Il faut noter que les personnes du cluster 3, représentent un peu plus de 15% du total des clients, et réalisent plus de 19% des retours, tandis que le cluster 5 possède 15% du total des clients, un peu plus de 8% du total des retours.

Tableau 4.8 Analyse de résultats – Tour de poitrine

	Cluster	Tendance	% Total	% Femmes	%Hommes	%Retours
1	Clus1		4,42%	5,05%	4,22%	1,69%
2	Clus2		57,93%	54,57%	58,98%	43,53%
3	Clus3		13,10%	14,51%	12,66%	13,51%
4	Clus4		12,20%	12,30%	12,17%	4,69%
5	Clus5		11,23%	12,93%	10,70%	6,19%
6	Clus6		1,12%	0,63%	1,28%	1,69%
7	Clus7		8,61%	13,88%	6,97%	28,71%
			100,00%	100,00%	100,00%	100,00%

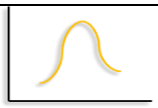
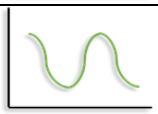
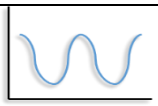
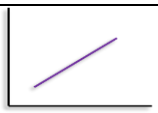
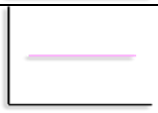
Ensuite, les résultats obtenus pour le tour de poitrine, selon le tableau 4.8, indiquent que plus de la moitié des clients se trouvent dans le cluster 2 (57,93%). Ce qui est intéressant dans cette analyse, c'est que le pourcentage de retours des clusters 1, 4 et 5 est faible par rapport au pourcentage de personnes que l'on retrouve dans chacun d'eux. D'un autre côté, les clients stables du cluster 7 qui représentent 8% du total des personnes font plus de 28% de retours.

Tableau 4.9 Analyse de résultats – Tour de taille

	Cluster	Tendance	%Clients	% Femmes	% Hommes	% Retours
1	Clus1		2,36%	2,17%	2,43%	2,19%
2	Clus2		3,61%	3,01%	3,84%	5,26%
3	Clus3		17,03%	17,56%	16,83%	22,22%
4	Clus4		25,36%	28,60%	24,12%	30,99%
5	Clus5		0,56%	0,17%	0,70%	0,44%
6	Clus6		21,15%	14,21%	23,80%	25,58%
7	Clus7		29,94%	34,28%	28,28%	13,30%
			100,00%	100,00%	100,00%	100,00%

Cette troisième analyse nous laisse voir 2 groupes qui incluent la plupart de clients (groupe 4 et 6). Néanmoins, la plupart des retours sont faits par les personnes des clusters 3 et 6. Nous pouvons observer une présence majoritaire des femmes dans le cluster 3. Le cluster des personnes stables présente également une quantité importante de femmes, mais on peut constater qu'en dépit d'être l'un des clusters avec le plus grand nombre de personnes, leur pourcentage de retours est faible.

Tableau 4.10 Analyse de résultats – Tour des hanches

	Cluster	Tendance	%Clients	% Femmes	%Hommes	% Retours
1	Clus1		36,07%	21,46%	27,88%	25,36%
2	Clus2		20,24%	17,88%	20,16%	18,53%
3	Clus3		3,35%	5,69%	2,43%	3,07%
4	Clus4		17,44%	19,19%	16,75%	18,70%
6	Clus5		3,49%	3,41%	3,52%	2,13%
7	Clus6		19,42%	32,36%	28,26%	32,19%
			100,00%	100,00%	100,00%	100,00%

Finalement, les derniers résultats (tableau 4.10) indiquent que la majorité des clients se trouvent dans le premier cluster (36,07%), où la tendance de la courbe monte et descend. Cependant, la plupart des retours ne sont pas faits par ce groupe, mais par les personnes du sixième cluster, où on trouve les clients stables. On souligne également la particularité des clusters 1, 3, 4 et 6 puisque la présence des femmes et majeure à celle des hommes.

Il faut se rappeler que dans l'analyse des résultats, il est possible qu'un client considéré comme stable dans l'une de ses mesures, puisse avoir un comportement variable, dans l'une des trois autres.

Les résultats obtenus à partir de cette segmentation, nous ont permis de trouver de manière satisfaisante des tendances, dans l'évolution des mensurations des clients. De plus, en joignant ces résultats aux informations de notre base de données, il a été possible de comprendre la composition de chacun des groupes que nous avons obtenus. Ainsi, il est possible d'identifier les groupes où se trouvent les clients qui réalisent le plus de retours, et le nombre d'hommes et de femmes présents dans chacun d'eux. Il est aussi important de souligner, que nos résultats de segmentation seront

utilisés comme l'une des variables d'entrée, pour notre modèle de prédiction qui sera développé à l'étape suivante.

4.5 Construction des arbres de décision

Les arbres de décision sont une méthode utilisée dans différentes disciplines comme modèle de prédiction. Au cours de cette étape, on va procéder au développement de notre modèle de prédiction à l'aide des arbres de décision.

Pour cette section, on a créé une nouvelle variable de type facteur avec deux niveaux : oui et non. Où « oui » signifie que le client effectue un retour, et « non » signifie que le client est satisfait de sa commande et ne la retourne pas. Dans le cas de l'arbre de décision pour le type de produits, « oui » signifie que le produit a été retourné et « non » que l'article n'a pas été retourné. Cette variable sera utilisée comme variable de sortie pour nos deux arbres de décision.

4.5.1 Choix de l'algorithme

Dans le domaine de l'apprentissage automatique, il existe différentes façons d'obtenir des arbres de décision. Le modèle que nous utiliserons est connu sous le nom de C5.0 (Kuhn et Johnson, 2013). Cet algorithme est successeur de C4.5 (Quinlan, 1993) dans le but de créer des arbres de classification. Parmi les caractéristiques de cette méthode, on note la capacité de générer des arbres de prédiction simples, des modèles basés sur des règles, des ensembles basés sur le boosting et l'attribution de différents poids aux erreurs se distinguent. Cet algorithme a été très utile pour créer des modèles de classification.

Pour sélectionner l'algorithme, nous avons effectué des tests initiaux avec un petit échantillon en utilisant les algorithmes ID3, C4.5 et C5.0. Ces 3 méthodes peuvent réaliser des arbres de classification en utilisant des variables qui ont plus de 50 facteurs. Cependant, contrairement à C4.5 et C5.0, le temps de calcul utilisé par ID3 était trop long. Le temps utilisé par C5.0 était inférieur au temps de C4.5. De plus, par rapport à son prédécesseur, l'algorithme C5.0 permet l'utilisation de techniques de « boosting » et « winnowing » pour améliorer la performance prédictive. Le «boosting» peut être utilisé avec n'importe quel nombre d'essais, plus d'essais donnent généralement des améliorations supplémentaires.

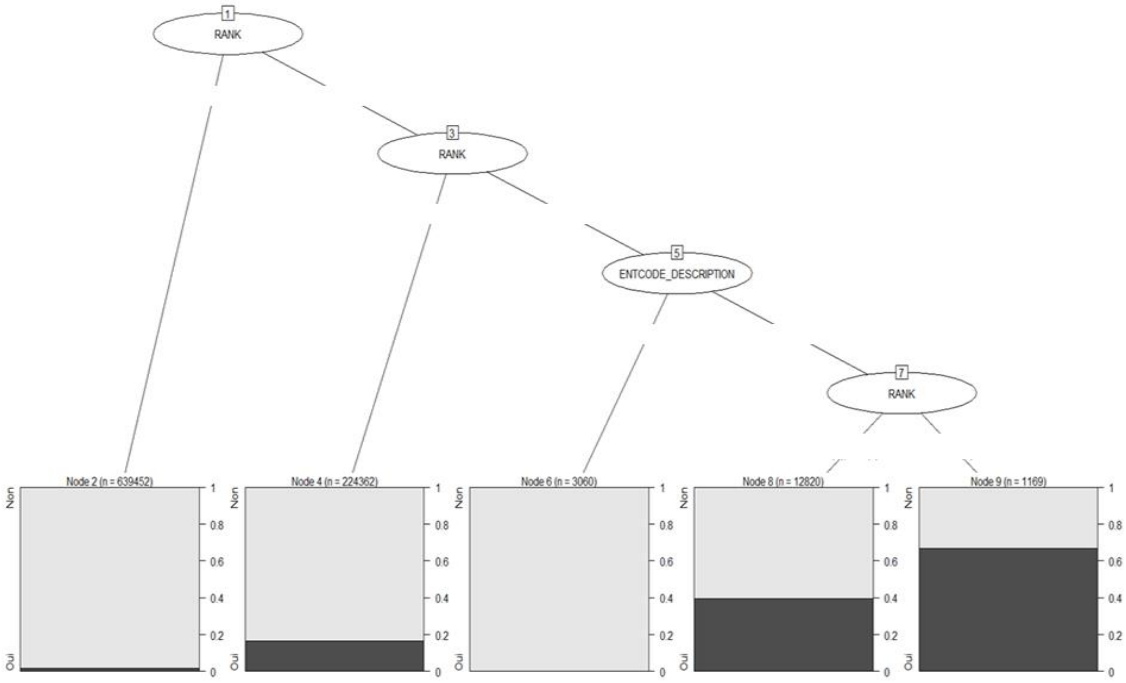


Figure 4.17 Exemple d'arbre de décision réalisé avec l'algorithme C5.0

La figure 4.17 nous montre un exemple réalisé avec un petit échantillon de notre base de données. Comme nous pouvons le voir, les nœuds d'extrémité nous montrent les résultats oui (couleur gris foncé) et non (couleur gris clair). Dans la partie supérieure de ceux-ci, le nombre total d'individus ou d'articles de chaque nœud est observé.

4.5.2 Boosting

Le boosting est une approche d'apprentissage automatique basée sur l'idée de créer une règle de prédiction très précise en combinant de nombreuses règles relativement faibles et imprécises. Pour chaque arbre, des tests ont été effectués en prenant entre 15 et 20 itérations, afin de choisir le résultat avec le pourcentage d'erreur le plus faible, et le taux de rendement le plus élevé. Dans le cas présent, nous avons utilisé des populations d'arbres simples.

Les résultats montrent que, grâce au processus de boosting, la capacité prédictive du modèle est améliorée puisque la performance a augmenté de 90,04% à 92,20% pour le premier arbre.

Tableau 4.11 Matrices de confusion avant et après le boosting – Arbre1

Avant	Prédiction	Non	Oui	Après	Prédiction	Non	Oui
Réel	Non	299341	37695	Réel	Non	495394	33793
	Oui	20248	242957		Oui	12968	58086

Le deuxième arbre a également montré une légère amélioration des performances, passant de 89,96% à 90,11%

Tableau 4.12 Matrice de confusion avant et après le boosting – Arbre 2

Avant	Prédiction	Non	Oui	Après	Prédiction	Non	Oui
Réel	Non	597188	44081	Réel	Non	606728	45321
	Oui	30317	69975		Oui	28917	69775

4.5.2.1 Identification des variables les plus influentes

C5.0 intègre deux façons de quantifier l'influence des prédicteurs sur le modèle final (simple ou ensemble).

Utilisation

C’est le pourcentage d'observations d'apprentissage qui tombent sur tous les nœuds générés, après une scission à laquelle le prédicteur a participé. Par exemple, dans le cas d'un arbre simple, le prédicteur utilisé dans la première décision reçoit automatiquement une importance de 100%, car toutes les observations de formation tombent sur l'un des deux feuilles générées. D'autres prédicteurs peuvent fréquemment participer aux divisions, mais si seules quelques observations tombent sur les nœuds enfants, leur importance peut être proche de zéro.

Partitionnements

C’est le pourcentage de divisions auxquelles chaque prédicteur participe. Il ne prend pas en compte l'importance de la division. Les deux mesures quantifient des aspects très différents, de sorte que le classement d'importance peut varier considérablement en fonction de celui qui est utilisé.

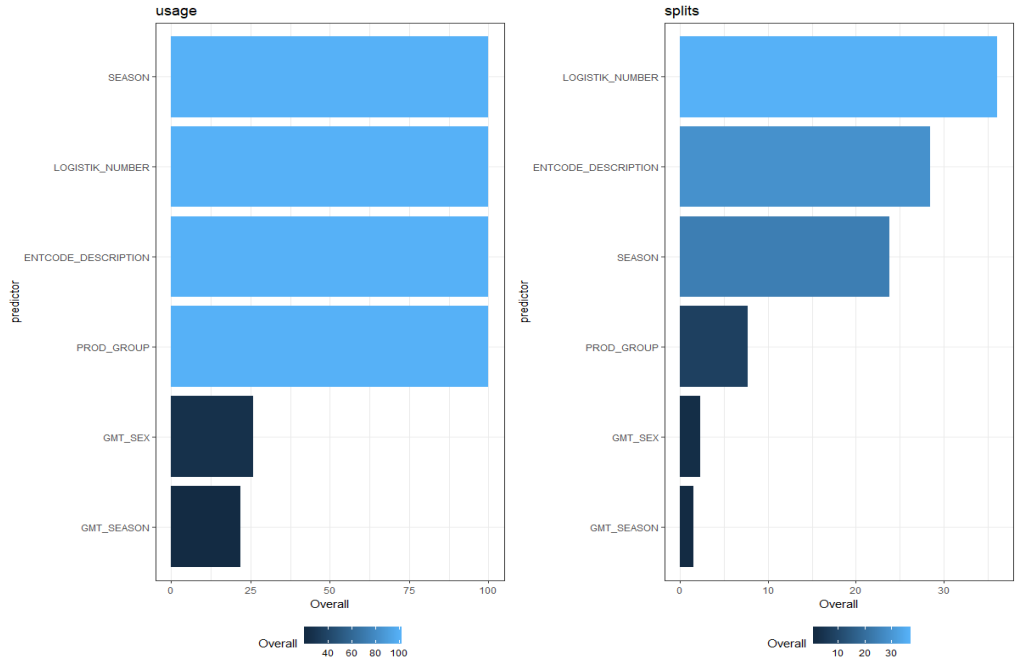


Figure 4.18 Variables influentes – Type de produits

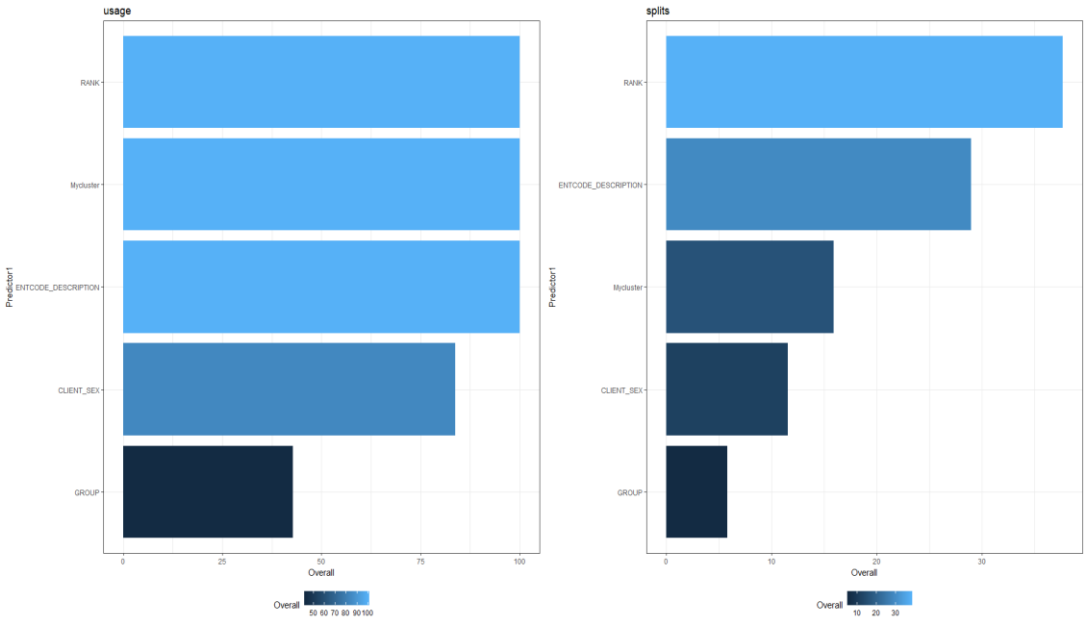


Figure 4.19 Variables influentes – Clients

Selon les résultats obtenus montrés dans les figures 4.18 et 4.19, les variables les plus influentes dans les deux arbres de décision, sont le groupe ou l'entité auquel la personne appartient (arbre de clients), et logiquement le type de produit dans notre premier arbre. En revanche, les variables qui indiquent le sexe dans les deux arbres ne présentent pas un niveau élevé de participation.

4.5.3 Winnowing

Le « winnowing » est une stratégie incorporée dans C5.0 qui permet de sélectionner des prédicteurs pertinents. Il consiste à utiliser un algorithme initial qui identifie les prédicteurs liés à la variable de sortie, puis on ajuste le modèle final uniquement avec ces prédicteurs. L'ensemble d'entraînement est divisé au hasard en deux moitiés. Avec l'un d'eux est ajusté un arbre de winnowing, dont le but est d'évaluer l'utilité des prédicteurs. Les prédicteurs qui n'ont été utilisés dans aucune division de l'arbre de winnowing sont identifiés comme inutiles. L'autre moitié des observations d'entraînement sont utilisées pour estimer l'erreur de l'arbre. Un processus itératif est lancé dans lequel chacun des prédicteurs, qui composent l'arbre de winnowing est éliminé (un par un). Si, à la suite de l'élimination, l'erreur d'arbre diminue, le prédicteur est considéré comme contribuant uniquement au bruit et, est identifié comme non utile. Une fois que tous les prédicteurs ont été identifiés comme utiles ou non utiles, l'arbre est réajusté en utilisant uniquement les prédicteurs utiles. Grâce au boosting effectué et à l'identification des variables importantes, lors de cette dernière étape, l'algorithme n'a éliminé aucune des variables, puisque l'erreur n'a pas diminué.

4.5.4 Type de vêtements

Notre premier modèle vise à développer un arbre de classification capable de prédire, si un vêtement avec certaines caractéristiques sera retourné ou non. Pour mener à bien cette étape, il a fallu préalablement sélectionner les variables jugées pertinentes pour construire le modèle.

On a pris différentes tailles d'échantillons pour les tests. On a commencé par prendre un petit échantillon d'un peu plus de 5000 articles. Pour construire des arbres de décision, il est nécessaire de diviser la base de données en deux parties, une base d'entraînement (70% du total dans ce cas) et une base de test (30% restant de la base) pour évaluer les performances de l'algorithme.

Plusieurs tentatives ont été faites pour le premier arbre de décision, afin de voir si les performances pourraient s'améliorer en changeant la taille d'échantillon. On a également décidé de tester si en supprimant l'une des variables d'entrée, les résultats varieraient. En supprimant des variables telles que le rang et l'entité, les résultats ont été considérablement affectés, puisque le taux de performance des arbres a été réduit entre 20% et 30%. Une autre variante essayée a été de filtrer la

base de données utilisée pour sélectionner des produits spécifiques ; cependant, les résultats n'ont pas beaucoup changé, le taux d'erreur restait entre 11% et 13%.

Enfin, nous sélectionnons un modèle en tenant compte du taux d'erreur, et du taux de performance (après les processus de boosting et winnowing). Le résultat de cet arbre, indique que les produits qui ont plus tendance à être retournés, ont les caractéristiques suivantes (les noms originaux des données ont été changés pour des raisons de confidentialité):

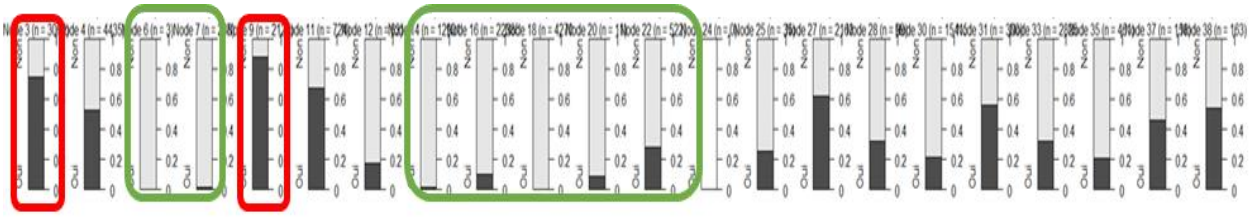


Figure 4.20 Résultats de l'arbre de décision – type de vêtements

La figure 4.20 nous montre les résultats du premier arbre de classification des produits, en tenant compte des variables d'entité, de grade et de groupe. Ainsi, on obtient que les produits les plus retournés sont les suivants :

- Les vêtements identifiés avec les numéros 7, 16, 17, 23, 165, 1014, 2700, 3657, 3660 fabriqués principalement après 2014.
- Les vêtements fabriqués pour les entités A2N, RF-2C, RF-A, RF-A-C, RF-N, RF-N-C et RFA2, avec les numéros 221, 245, 266, 286, 1145 et 1145.
- Les vêtements appartenant aux groupes 1, 2, 13, 14, 15, 18, 20 et 21.
- Les vêtements identifiés avec les numéros 1031 fabriqués après la saison 2015
- Les vêtements identifiés avec les numéros 44, 143, 147, 174, 175, 233, 1031 et 1113 destinés aux entités RF-A
- Les vêtements des groupes 8, 10, 11 et 16 fabriqués pour les entités RF-A2-C et RF-A-C

Au contraire, les vêtements qui ont le plus faible pourcentage de retour sont les vêtements identifiés avec les numéros 43, 44, 56, 102, 131, 143, 147, 173, 174, 175 176, 233, 1062, 1113, 2507, 2508 2509 réalisées à partir de la saison 2015 pour les entités A2O, AN, AO, H2A, HA, HN et NN.

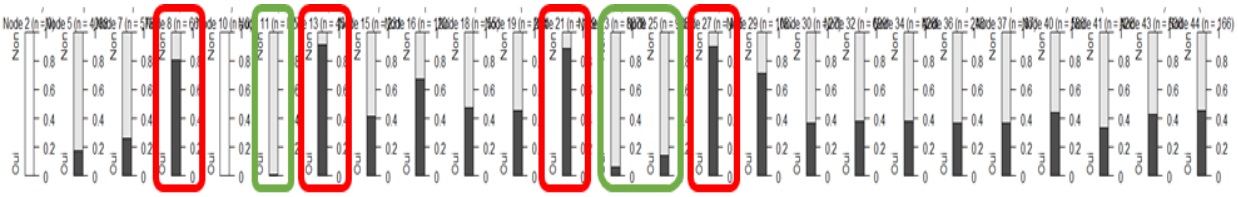


Figure 4.21 Arbre de décision – type de vêtements

La figure 4.21 nous montre aussi les résultats de l'arbre de classification réalisé pour les types de vêtements. Il est à noter que dans cet exemple, les variables d'entité et de groupe n'ont pas été incluses. Comme on peut le voir en rouge, 4 des nœuds d'extrémité présentent un nombre élevé de cas de retour, tandis que les nœuds encadrés en vert montrent les produits les moins retournés. Les vêtements présentant les caractéristiques suivantes ont une probabilité supérieure aux autres d'être retournés.

- Les vêtements des groupes 1, 2 ,6 ,7 ,8 ,11 ,12 ,13 ,14 ,15 ,16 et 20 (Parmi ces numéros, on trouve les chemises, les pantalons et les manteaux) des saisons 2004 - 2010
- Les produits 173, 174, 175, 177, 178, 221, 237, 240, 266, 1031, 1144 et 1145 faits pour l'hiver ou les deux saisons.
- Les produits 339 et 1136 des saisons 2011 – 2018
- Les produits 56, 102 et 245 des saisons 2015 – 2018

En revanche, les vêtements les moins retournés ont une caractéristique commune qui indique que ce sont principalement ceux d'été.

En comparant les résultats, on observe qu'il y a des groupes et des nombres de vêtements qui apparaissent dans les deux exemples, comme c'est le cas des groupes 1, 2, 13 et 14, les vêtements fabriqués à partir de 2015 et qui portent les numéros 174, 175, 221 et 1031.

4.5.5 Clients

Le modèle présenté ci-dessous a été développé afin de trouver les caractéristiques des clients qui ont tendance à effectuer des retours. Pour commencer, il faut sélectionner les variables d'entrée. Comme dans le premier arbre, la première étape a été de diviser la base en : ensemble d'apprentissage (70%) et de test (30%). Comme à l'étape précédente, nous avons décidé d'effectuer différents tests en modifiant les variables d'entrée et la taille de l'échantillon.

4.5.5.1 Ajout de groupes aux résultats

Dans cette phase, on ajoutera en tant que variable d'entrée les clusters créés dans la section précédente, puis on analysera si notre arbre montre une amélioration de son taux de performance. Initialement, nous n'avons considéré que la première mesure, c'est-à-dire les grappes obtenues pour le contour du cou ; en conséquence, nous avons obtenu une légère amélioration des performances de l'arbre, de 87,86% à 89,21% (en prédiction). Nous avons ensuite procédé à l'ajout des trois autres mesures. Cependant, les performances variaient très peu. Par conséquent, nous avons décidé d'utiliser chaque mesure séparément, c'est-à-dire dans 4 arbres différents, et de comparer les résultats. Ci-dessous, nous présentons les résultats les plus significatifs avec les meilleures performances (après le processus de boosting et de winnowing) pour chacune des mesures. Il convient de noter que les noms réels de certaines catégories, ont été modifiés pour des raisons de confidentialité.

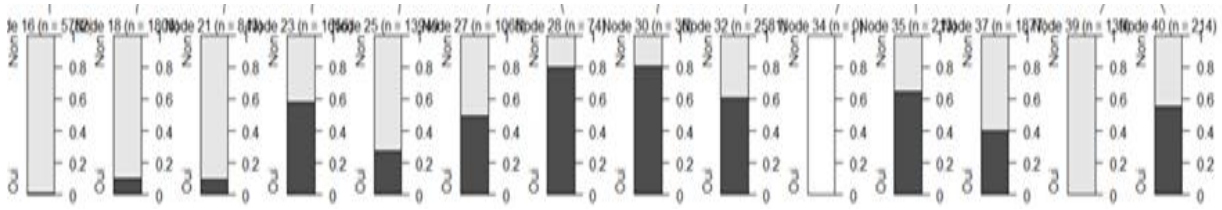


Figure 4.22 Arbre de décision – Mesure Tour de cou

La figure 4.22 nous permet de voir les résultats de notre premier arbre de classification des clients. Ainsi, on en conclut que les clients qui effectuent le plus grand nombre de retours présentent les caractéristiques suivantes :

- Hommes de l'entité RF-A, des grades B, R et PCLP, appartenant au cluster 3 des groupes C et A.
- Les femmes de l'entité RFA2, des grades B, T et CPL, et qui appartiennent également au cluster 4.
- Femmes des entités AN2, RF-A, RFA2, des groupes C et A appartenant aux clusters 2, 3, 4 et 5
- Hommes et femmes du groupe 7 du groupe AB, des rangs CA, LS et MC.

En revanche, les personnes qui n'effectuent jamais ou presque jamais de retours sont des clients des clusters 4 et 7 appartenant aux grades CWO, LO, CA, LS, OCDT et PO2.

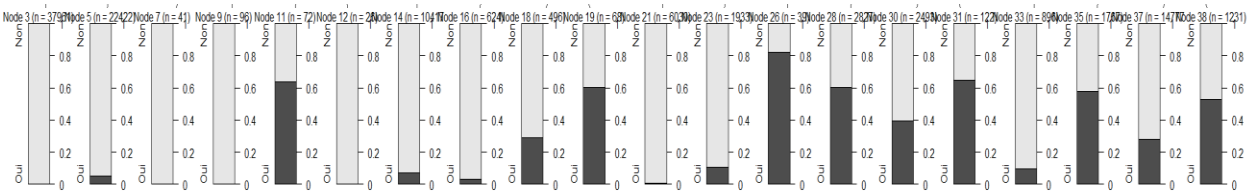


Figure 4.23 Arbre de décision – Mesure Tour de poitrine

Les résultats obtenus pour l'arbre de décision de la mesure « tour de poitrine » sont présentés sur la figure 4.23. Ainsi, on en conclut que les personnes qui font des retours ont plus souvent les caractéristiques suivantes :

- Hommes des codes 261 et 337
- Hommes des codes 10, 71, 339 et 337 du groupe 4
- Femmes appartenant aux clusters 2, 4 et 5 de l'entité RF - A, grade CPL avec les codes 171 et 376
- Les femmes des entités RF-A et RFA2, du grade 2LT des clusters 2 et 5.
- Hommes de l'entité RF-A, rang 2LT, groupe A appartenant aux clusters 2 et 5.
- Hommes appartenant aux clusters 1 et 3 des entités RF-A, RF-A-C et RF-N, des grades CA et LS
- Femmes des clusters 1 et 3, appartenant à l'entité RF-N
- Femmes des clusters 1 et 6 du groupe R, code 129

De même, il est possible d'observer que les personnes des entités A2R, AN, AR, AO, NN, NO et R-A-C du cluster 7 sont celles qui font le moins de retours.

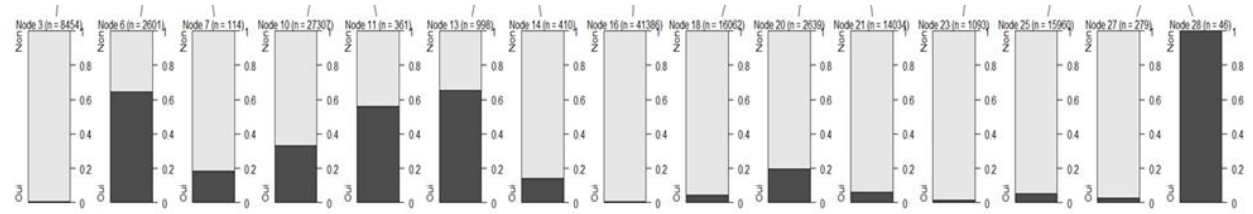


Figure 4.24 Arbre de décision – Mesure Tour de taille

Dans la figure 4.24, les résultats du troisième arbre de classification (3) sont affichés. Les clients qui ont le plus tendance à effectuer des retours sont:

- Hommes et femmes du groupe 6 avec les rangs 2LT, LT, OCDT
- Femmes et hommes appartenant aux clusters 1, 2, 4 et 5 des entités RF-A et RFA des grades B R et T

- Femmes et hommes du grade CPL, groupe RA des clusters 6 et 7
- Femmes de l'entité RFA, groupe R des clusters 4 et 7.

D'un autre côté, les clients qui font le moins de retours sont ceux appartenant à l'entité RF-A-C des groupes C, R et B.

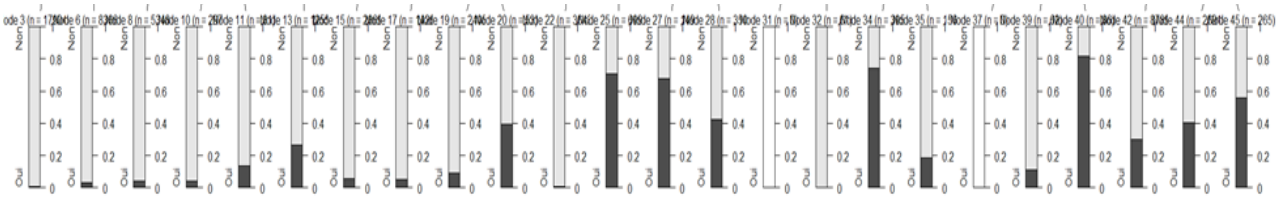


Figure 4.25 Arbre de décision – Mesure Tour des hanches

La figure 4.25 montre l'arbre résultant pour la mesure "tour des hanches". Les résultats indiquent que les clients qui ont tendance à faire le plus de retours présentent les caractéristiques suivantes:

- Hommes des clusters 1 et 4 appartenant aux grades CA et MC
- Femmes et hommes des clusters 2, 3, 4 et 6 appartenant à l'entité RF-A et au groupe RA.
- Femmes des clusters 1, 3, 4 et 5 appartenant à l'entité RFA et de rang CPL
- Femmes et hommes du grade T appartenant aux clusters 2 et 6
- Femmes et hommes appartenant aux clusters 1, 2 et 6 des grades B et R

En revanche, les personnes qui font le moins de retours sont:

- Clients de l'entité RF-A-C
- Femmes et hommes de rang T appartenant au cluster 1

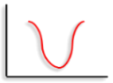
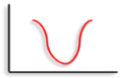
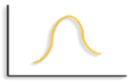
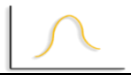
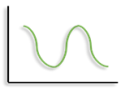
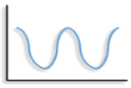
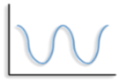
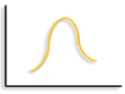
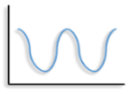
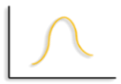
Après avoir effectué l'analyse comparative entre les 4 arbres, on constate que même si les comportements les plus susceptibles de générer des rendements varient d'une mesure à l'autre, il existe des variables telles que l'entité, l'intervalle et le groupe qui se répètent dans les 4 cas. En effet, les individus des entités RFA, RF-A et RA, avec le grade CPL, des groupes B, T et R ont tendance à effectuer un grand nombre de retours plus souvent que les autres clients.

4.6 Conclusion

Au cours de ce chapitre, on a commencé par présenter notre partenaire industriel. On a procédé à l'application de la méthodologie exposée dans le chapitre précédent, en commençant ce processus par la préparation de nos données pour connaître la situation actuelle du partenaire industriel.

Ensuite, nous avons procédé à la conversion de nos données en séries temporelles afin de mieux visualiser l'évolution des mensurations des clients. Quatre séries temporelles ont été construites pour chaque client, chacune d'entre elles représente l'évolution des mensurations principales. Nous avons continué par chercher des profils d'évolutions similaires en utilisant une méthode de segmentation, après cette étape nous avons regroupé les comportements similaires en nouveaux clusters. Chaque nouveau cluster a été analysé afin de connaître la quantité d'hommes et de femmes ainsi que le pourcentage de retours effectués. Finalement, nous avons réalisé des arbres de décision pour trouver les caractéristiques des clients qui font des retours et les caractéristiques des produits qui sont le plus souvent retournés. Les approches de winnowing et de boosting nous ont permis d'améliorer la performance de chaque arbre. Grace à ces résultats, nous avons trouvé que les personnes classées comme des clients qui font des retours sont généralement les individus des entités RFA, RF-A et RA avec le grade CPL, des groupes B, T et R. De la même manière, nous avons trouvé que les produits problématiques (plus souvent retournés) sont principalement les vêtements des groupes 1, 2 13, 14 et 15, faits pour l'hiver ou bi saison et qui ont été fabriqués après la saison 2014. La segmentation et les arbres des décision nous ont aussi permis de noter l'évolution des mensurations des clients qui font plus des retours. Les résultats ont été groupés sur le tableau 4.13.

Tableau 4.13 Comportements des personnes qui font la plupart des retours

Mesure	Hommes	Femmes
Tour de cou		 
Tour de poitrine	 	
Tour de taille	 	  
Tour des hanches	 	  

Il est possible de visualiser des profils similaires sur le tableau 4.13, par exemple, pour la mesure « tour de cou » les hommes et les femmes avec une tendance qui commence par descendre mais qui remonte à la fin sont classées comme des personnes qui font des retours. Il est aussi possible de voir quelques tendances qui se répètent pour plus d'une mesure. Par exemple, on trouve la tendance 6 (courbe qui monte et qui descends à la fin) pour « le tour de cou », « le tour de poitrine » et « le tour des hanches ».

CHAPITRE 5 CONCLUSIONS, PERSPECTIVES ET RECOMMANDATIONS

Comme mentionné tout au long de ce projet, les retours de produits sont toujours considérés comme un problème coûteux, en particulier pour les entreprises de commerce électronique. L'objectif principal de ce projet était de développer un modèle ou un outil permettant à notre partenaire de reconnaître les caractéristiques des individus qui effectuent des retours, afin d'anticiper et d'éviter un éventuel retour futur. Bien que dans la littérature actuelle, il existe de nombreuses études et applications (mentionnées au chapitre 2) pour la prédiction des retours. La plupart de ces travaux visaient à connaître le nombre d'articles ou de produits qui seraient retournés. Cependant, comme notre partenaire se consacre à la fabrication de vêtements personnalisés, un seul produit, comme une chemise, pourrait avoir 100 variations en raison des préférences individuelles de chaque client. En ce sens, il a été décidé de procéder à la recherche des caractéristiques des clients qui effectuent des retours. Pour cela nous avons développé un modèle de prédiction capable d'identifier les profils de personnes qui font des retours plus souvent et les caractéristiques des produits retournés.

Des données discrètes de commandes ont été modélisées pour être traitées avec des algorithmes d'analyse de comportements continus. En utilisant des algorithmes d'apprentissage non supervisé (segmentation) et supervisé (arbres de décision), nous avons pu trouver les clients et les produits les plus fréquemment retournés. Pour effectuer la segmentation, il a d'abord fallu construire des séries chronologiques à partir des commandes passées par les clients pendant toute la période d'étude. Nous avons choisi quatre mesures principales, à partir desquelles nous avons procédé à la construction de nos séries. Ainsi, nous avons été confrontés à un premier obstacle dû au caractère intermittent de nos séries temporelles. Pour cette raison, nous nous sommes appuyés sur une étude réalisée précédemment au sein de l'entreprise partenaire (El Ayeb et Agard, 2019) pour choisir la granularité la plus adaptée à nos séries, ainsi que la méthode d'agrégation appropriée. En utilisant la segmentation nous avons divisé l'ensemble des clients selon leurs mensurations, nous avons ensuite utilisé ces mesures afin de créer une nouvelle variable d'entrée pour le modèle de précision. Grâce à celle-ci, nous avons été capables de reconnaître les tendances des clients qui effectuent le plus grand nombre de retours. De plus, nous avons également pu observer les pourcentages, ainsi que les caractéristiques de chacun des clusters, comme le nombre de femmes et d'hommes qui les composent.

Les résultats de notre segmentation ont montré que l'évolution des mesures d'une personne peut suivre différentes tendances, c'est-à-dire qu'une personne qui augmente par exemple sa taille de tour de poitrine, ne va pas nécessairement augmenter ses autres mesures. Une autre observation importante a montré que même les clients avec une tendance linéaire (c'est-à-dire qui n'ont jamais changé de taille) effectuent un nombre considérable de retours. Ce fait pourrait impliquer des erreurs d'expédition ou un mécontentement quant à la qualité des produits reçus.

La construction d'arbres de classification est une alternative de classification / discrimination traditionnelle à la prédiction traditionnelle. Pour notre analyse, 2 arbres ont été créés; un pour les clients et un pour les vêtements. Nous avons choisi l'algorithme C5.0 et nous avons utilisé les approches de boosting et de winnowing afin d'obtenir des meilleures performances. L'arbre développé pour les types de vêtements avait pour objectif de faire connaître les caractéristiques des produits retournés, c'est-à-dire que grâce à ce modèle, nous sommes en mesure de savoir si par exemple un pantalon d'hiver vert de la saison 2015 a un pourcentage de rendement plus élevé qu'une chemise à manches courtes pour deux saisons (hiver-été). Les résultats ont montré que la plupart des vêtements retournés sont des vêtements deux saisons et des vêtements d'hiver fabriqués à partir de la saison 2014.

D'autre part, l'arbre de classification conçu pour les clients était destiné à reconnaître les caractéristiques des personnes qui effectuent des retours. Quatre arbres ont été construits, un pour chaque mesure. Nous avons commencé par utiliser les variables jugées pertinentes de notre base de données, puis nous avons procédé à l'ajout des clusters que nous avons obtenus à l'étape précédente. Les résultats de ce deuxième arbre nous ont permis d'identifier les entités, les grades et les groupes auxquels appartiennent les clients qui effectuent des retours plus souvent, ainsi que leurs comportements (la variation de leurs mesures). Ces résultats permettront à notre partenaire de prendre les mesures nécessaires à chaque fois qu'un client classé comme susceptible de faire un retour passe une nouvelle commande, surtout si cette commande comporte l'un des vêtements ayant le taux de retour le plus élevé.

Bien que les arbres de décision nous aient permis d'obtenir les résultats attendus, il existe d'autres techniques et algorithmes, comme les réseaux de neurones, qui pourraient fournir des résultats plus précis. C'est pourquoi, si une étude plus approfondie est menée, l'utilisation de ces autres techniques est jugée souhaitable.

Ce modèle de prédiction peut être reproduit et utilisé pour des entreprises qui ont un historique des commandes de leurs clients, dans notre cas, pour notre partenaire Logistik il est utilisé afin de pouvoir permettre à l'entreprise de prendre des mesures préventives avec les personnes qui ont tendance à faires des retours, surtout s'ils commandent des produits qui sont souvent retournés.

On recommande à notre partenaire de demander aux clients de prendre leurs mesures à chaque fois qu'ils passent une nouvelle commande, même si le délai entre leur dernière et leur nouvelle commande est court. Car pendant cette étude, au cours de la construction des séries chronologiques il a été constaté que certains clients présentent des variations excessives (4 ou plus) sur de courtes périodes (moins de 6 mois), ce qui indiquerait que les clients ne mettent pas régulièrement à jour leurs mesures.

Une réévaluation des mesures utilisées, à la fois par notre partenaire et ses clients est aussi considérée comme importante, puisqu'il est possible que la façon dont les clients prennent leurs mesures ne soit pas la même que celle de notre partenaire. Finalement, il est également conseillé de vérifier le processus de fabrication des vêtements, en particulier ceux qui ont un pourcentage élevé de retours, car il pourrait y avoir un problème avec les machines ou le personnel.

RÉFÉRENCES

Abe, M., et Nakayama, H. (2018). Deep learning for forecasting stock returns in the cross-section. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Juin, pp. 273-284. Springer, Cham.

Adams, N., Marquez, D. et Wakefield, G. (2005). Iterative deepening for melody alignment and retrieval. *In Proceedings of International Society for Music Information Retrieval*, pp. 199–206.

Adwan, S. et Arof, H. (2012). On improving Dynamic Time Warping for pattern matching. *Measurement*, vol. 45, n°6, pp. 1609-1620.

Aghabozorgi, S., Shirkhorshidi, A.S. et Wah T.Y. (2015) Time-series clustering a decade review, *Information Systems*, vol. 53, pp. 16–38.

Arrechea, J. E. (2011). *Vender en Internet: Las claves del éxito*, ed. Anaya Multimedia, Madrid, Espagne.

Au, T. C. (2018). Random forests, decision trees, and categorical predictors: the "Absent levels" problem. *The Journal of Machine Learning Research*, vol.19, n°1, pp. 1737–1766.

Ayers, J. B. (2001). *Handbook of Supply Chain Management.* Boca Raton. Fla.: The St. Lucie Press/APICS Series on Resource Management, New York, pp. 488.

Bali, T. G., et Hovakimian, A. (2009). Volatility spreads and expected stock returns. *Management Science*, vol. 55, n° 11, pp. 1797-1812.

Ballesteros Riveros, D. et Ballesteros Silva, P. (2008). Importancia de la administración logística. *Scientia Et Technica*, vol. 1, n° 38.

Ballou, R. H. (2004). *Logística: Administración de la cadena de suministro*. Pearson Educación, Mexique.

Begum, N., Ulanova, L., Wang, J. et Keogh, E. (2015). Accelerating dynamic time warping clustering with a novel admissible pruning strategy. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Août, pp. 49-58.

Berkhin, P. (2006). A Survey Of Clustering Data Mining Techniques. *In Grouping Multidimensional Data: Recent Advances in Clustering*, Springer, Berlin, Heidelberg, pp. 25-71.

Bernon, M., Cullen, J. et Gors, J. (2016). Online retail returns management: Integration within an omni-channel distribution context. *International Journal of Physical Distribution & Logistics Management*, vol. 46, n° 584.

Bolduc, C. (2014). *Commerce électronique : Les Canadiens achètent 51% de leurs produits au Canada*, tiré de: https://isarta.com/infos/le-commerce-electronique-les-canadiens-achetent-51-de-leurs-produits-au-canada/.

Breiman, L., Friedman, J. H., Olshen, R. A. et Stone, C. J. (1984). Classification and Regression Trees, *Statistics/Probability*, Wadsworth, California, USA.

Bruzzone, A., Cimino, A., Diaz, R. et Longo, F. (2010). Inventory control with products returns: a state of the art overview. *Paper presented at the Proceedings of the 2010 Spring Simulation Multiconference*, Avril, Orlando, Florida, pp 1- 7.

Carrión García, A. (2001). Análisis de series temporales, técnicas de previsión. *Publicaciones de la Universidad Politécnica de Valencia*, n°795.

Carter, J, (2018). *Forbes B2B E- commerce Trends to take Notice of in 2018*, tiré de: https://www.forbes.com/sites/theyec/2018/02/15/b2b-e-commerce-trends-totake-notice-of-in-2018.

Canadapost (2020). *The 2020 Canadian e-commerce report*, tiré de: https://www.canadapost.ca/cpc/en/business/marketing/campaign/ecommerce-report.page.

Canda, A., Yuan, X. M. et Wang, F. Y. (2015). Modeling and forecasting product returns: An industry case study. *In 2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Décembre 6-9, Singapore, pp. 871 – 875.

Cao, F., Ester, M., Qian, W. et Zhou, A. (2006). Density-based clustering over an evolving data stream with noise. I*n Proceedings of the 2006 SIAM international conference on data mining*, Society for industrial and applied mathematics, pp. 328-339

Cefrio, (2017). *Netendances 2018 - le commerce électronique au Québec*, tiré de: https://cefrio.qc.ca/fr/enquetes-et-donnees/netendances2018-commerce-electronique-au-quebec, 2019.

Cendejas, J. L. (2016). Introducción a las series temporales I. Mémoire de maîtrise, Universidad Francisco de Vitoria, Madrid, Espagne.

Chatfield, C. et Xing, H. (2019). *The analysis of time series: an introduction with R*. CRC press, pp. 15-37.

Chitrakorn, K. (2018). *5 Technologies Transforming Retail in 2018*, tiré de: https://www.businessoffashion.com/articles/technology/5-technologies-transforming-retail.

Choi, T.M., Li, D. et Yan, H. (2004). Optimal returns policy for supply chain with e-marketplace. *International Journal of Production Economics*, vol. 88, n° 2, pp. 205-227.

Closs D.J. et Savitskie K., (2003). Internal and External Logistics Information Technology Integration. *The International Journal of Logistics Management*, vol. 14, n° 1, pp.63-76.

Croston, J.D. (1972), Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, vol. 23, n°3, pp. 289-303.

Cryer, J. D. et Chang, K.S (2008). *Time series analysis: with applications in R*. Springer Science & Business Media, pp. 1-15.

De Brito, M., Flapper, S. D., et Dekker, R. (2002). *Reverse logistics*. No. EI 2002-21.

De Brito, M. P. et van der Laan, E. A. (2009). Inventory control with product returns: The impact of imperfect information. *European Journal of Operational Research*, vol. 194, n° 1, pp. 85-101.

Del Mondo G., (2011) A Spatio-temporal graph-based model for the evolution of geographical entities, Thèse de doctorat, Université de Bretagne Occidentale, Brest, France.

De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews Computational Statistics*, vol. 5, n° 6, pp. 448-455.

Dekker, R., Fleischmann, M., Inderfurth, K., et van Wassenhove, L. N. (2013). *Reverse logistics: quantitative models for closed-loop supply chains.* Springer Science & Business Media, pp. 3 – 48.

Deng, S., Li, Y., Guo, H., et Liu, B. (2016). Solving a Closed-Loop Location-Inventory-Routing Problem with Mixed Quality Defects Returns in E-Commerce by Hybrid Ant Colony Optimization Algorithm. *Discrete Dynamics in Nature and Society*, 6467812.

Dennis, S. 2018. *The Ticking Time Bomb Of Ecommerce Return*, tiré de: https://www.forbes.com/sites/stevendennis/2018/02/14/the-ticking-time-bombof-e-commerce-returns/#58b30eb24c7f.

Donaldson, T. (2015). *E-commerce return rates expected to exceed 30%.* tiré de : https://sourc ingjo urnal .com/topic s/retai l/e-commerce-return-rates-expected-to-exceed-30-39222 /.

Enke, D. et Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, vol. 29 n° 4, pp. 927-940.

El Ayeb S. et Agard B., (2019) Développement d'un outil de segmentation des comportements d'achat des clients en se basant sur leurs données morphologiques, Mémoire de maîtrise, École Polytechnique de Montréal, Montréal, Canada.

Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A. et Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, vol. 41, n°4, Part 2, pp. 1937-1946.

Ferreira, M. A. et Santa-Clara, P. (2011). Forecasting stock market returns: The sum of the parts is more than the whole. *Journal of Financial Economics*, vol. 100, n° 3, pp. 514-537.

Ferrell, L., Hirt, G. A. et Ferrell, O. C. (2010). *Introducción a los negocios en un mundo cambiante*. Ed. Mc Graw Hill, México D.F, Mexique.

Ferri, F., Grifoni, P. et Guzzo, T. (2013). Factors determining mobile shopping. A theoretical model of mobile commerce acceptance. *International Journal of Information Processing and Management*, vol. 4, n° 7, pp. 89-101.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, vol. 21, pp. 768-769.

Forrester Consulting (2008). *Crafting a Returns Policy that creates a Competitive Advantage Online, A commissioned study conducted by Forrester Consulting on behalf of UPS.* Forrester: Cambridge, Massachusetts, USA.

Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, vol. 24, n° 1, 164-181.

Fu, C., Zhang, P., Jiang, J., Yang, K. et Lv, Z. (2017). A Bayesian approach for sleep and wake classification based on dynamic time warping method. *Multimedia Tools and Applications*, vol. 76, n° 17, pp. 17765-17784.

Garicano, L. et Kaplan, S. N. (2001). The Effects of Business-to-Business E-Commerce on Transaction Costs. *The Journal of Industrial Economics*, vol. 49, n° 4, pp. 463-485.

Gary Schneider, P. (2006*) Electronic Commerce*, 6th ed., Thomson Learning, Beijing, pp.5-8.

Gehrke, J., Loh, W.-Y. et Ramakrishnan, R. (1999). Classification and regression: money can grow on trees. Tutorial notes for ACM SIGKDD 1999 *International conference on Knowledge Discovery and Data Mining*, Août 15-18, California, USA, pp. 1-73.

Geler, Z., Kurbalija, V., Ivanović, M., Radovanović, M. et Dai, W. (2019). Dynamic Time Warping: Itakura vs Sakoe-Chiba. *Paper presented at the 2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, Juillet 3-5, pp. 1-6.

Giner, I. G., Alvarez, A. R., Sánchez-Cambronero García-Moreno, S. et Camacho, J. L. (2016). Dynamic modelling of high speed ballasted railway tracks: analysis of the behaviour. *Transportation Research Procedia*, vol. 18, pp. 357-365.

Gold, O. et Sharir, M. (2018). Dynamic Time Warping and Geometric Edit Distance: Breaking the Quadratic Barrier. ACM Transactions on Graphics. *Algorithms*, vol. 14, n°4.

García González, R. (2014). *El comercio electrónico en España.* Repositorio de Universidad de Cantabria, n°10902.

Graf, A., et Schneider, H. (2016). *The E-Commerce Book: About a channel that became an industry*. Mediengruppe Fachbuch.

Griffis, S.E., Rao, S., Goldsby T.J. et Niranjan, T.Y. (2012). The customer consequences of returns in online retailing: An empirical analysis, *Journal of Operations and Management*, vol. 30, n°. 4, pp. 282–294.

Guide, J. V. et Van Wassenhove, L. N. (2002). The reverse supply chain. *Harvard business review*, vol. 80, n° 2, pp. 25-26.

Gudmundsson, G. (1971). Time-Series Analysis of Imports, Exports and Other Economic Variables. *Journal of the Royal Statistical Society: Series A (General)*, vol. 134, n° 3, pp. 383-412.

Gunasekaran, A., Patel, C. et Tirtiroglu, E. (2001). Performance measures and metrics in a supply chain environment. *International Journal of operations & production Management*, vol 21, pp-71-87.

Gutierrez, V. et Vidal, C. (2008). Modelos de gestión de inventarios en cadenas de abastecimiento: Revisión de la literatura. *Revista de la Facultad de Ingeniería de la Universidad de Antioquia*, Colombia, n° 43, pp. 134-149.

Han, J. (2005). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc. pp.5-21.

Han, J., Kamber, M. et Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc, pp. 8-32.

Hao, Y. (2014). *Returns Reverse Logistics Management Strategy in E-commerce B2C Market*.

Hofacker, L. et Langenberg, C. (2015). *E-Commerce-Markt Osterreich/Schweiz*. EHI Retail Institute.

Hjort, K. et Lantz, B. (2016). The impact of returns policies on profitability: A fashion e-commerce case. *Journal of Business Research*, vol. 69, n° 11, pp. 4980-4985.

Hssina, B., Merbouha, A., Ezzikouri, H. et Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. (IJACSA*) International Journal of Advanced Computer Science and Applications*, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, vol 4, n°2, pp. 13-19.

Huang, S. H. (2013). *Supply Chain Management for Engineers 1st Edition ed*., pp. 240.

Huang, X., Ye, Y., Xiong, L., Lau, R. Y. K., Jiang, N. et Wang, S. (2016). Time series k-means: A new k-means type smooth subspace clustering for time series data. *Information Sciences*, vol. 367-368, pp. 1-13.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, vol. 23, n°1, 67-72.

Izakian, H., Pedrycz, W. et Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, vol. 39, pp. 235-244.

Janakiraman, N., Syrdal, H. A. et Freling, R. (2016). The Effect of Return Policy Leniency on Consumer Purchase and Return Decisions: A Meta-analytic Review, *Journal of Retail*, vol. 92, n° 2, pp. 226–235.

Janakiraman, G. et Seshadri, S. (2017). Dual Sourcing Inventory Systems: On Optimal Policies and the Value of Costless Returns. *Production and Operations Management*, vol. 26, n° 2, pp. 203-210.

Jeong, Y.-S., Jeong, M. K. et Omitaomu, O. A. (2011). Weighted dynamic time warping for time series classification. *Pattern Recognition*, vol. 44, n° 9, pp. 2231-2240.

Jeszka, A. M. (2014). Product Returns management in the clothing industry in Poland. *Logistics Forum*, vol. 10, n° 4, pp. 433-443.

Johnson, C. (2018). *US eCommerce Forecast: 2008 To 2012*, tiré de: https://www.forrester.com/report/US+eCommerce+Forecast+2008+To+2012/-/E-RES41592#.

Joshi, T., Mukherjee, A. et Ippadi, G. (2018). One Size Does Not Fit All: Predicting Product Returns in E-Commerce Platforms. *Paper presented at the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* Août 28-31, pp. 926-927.

Jullien, M. (2019). Analyse temporelle de la diversité en régime autogame : Approches théorique et empirique. Thése de Doctorat, Montpellier SupAgro, Montpellier, France.

Kantavat, P., Kijsirikul, B., Songsiri, P., Fukui, K.-I. et Numao, M. (2018). Efficient Decision Trees for Multi–Class Support Vector Machines Using Entropy and Generalization Error Estimation. *International Journal of Applied Mathematics and Computer Science*, vol. 28, n° 4, pp. 705.

Kaufman, L. et Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons.

Keogh, E. J. et Pazzani, M. J. (2001). Derivative dynamic time warping. *In Proceedings of the 2001 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics. Avril, pp. 1-11.

Kim, J. H., Shamsuddin, A. et Lim, K.-P. (2011). Stock return predictability and the adaptive markets hypothesis: Evidence from century-long U.S. data. *Journal of Empirical Finance*, vol. 18, n° 5, pp. 868-879.

Kirchgässner, G., Wolters, J. et Hassler, U. (2007*). Introduction to Modern Time Series Analysis*. Berlin: Springer, pp 1-22.

Kokkinaki, A. I., Dekker, R., Koster, M. B. M. d., Pappis, C. et Verbeke, W. (2002). E-business models for reverse logistics: contributions and challenges. *Proceedings International Conference on Information Technology: Coding and Computing*. Avril, 8-10, Las Vegas, États-Unis, pp 470-476.

Kotler, P. (2003). *Marketing management 11th ed*., Prentice Hall of India, New Delphi, pp. 26-27.

Kristensen, K., Borum, N., Christensen, L. G., Jepsen, H. W., Lam, J., Brooks, A. L. et Brooks, E. P. (2013). Towards a next generation universally accessible ‘online shopping-for-apparel’system. *In International Conference on Human-Computer Interaction*, Juillet, Springer, Berlin, Heidelberg, pp. 418-427.

Kourentzes, N. (2013). Intermittent demand forecasts with neural networks. *International Journal of Production Economics*, vol. 143, n° 1, pp. 198-206.

Kuhn, M. et Johnson, K. (2013). *Applied predictive modeling*, vol. 26. New York, Springer, pp 159-207.

Kuhn, M., Weston, S., Coulter, N. et Culp, M. (2015), *C50: C5.0 decision trees and rule-based models*, tire de: https:// CRAN.R-project.org/package=C50, R package version 0.1.0-24, C code for C5.0 by R. Quinlan.

Lalonde, D. (2017). *Les Québécois ont dépensé 8,5 milliards $ en ligne en 2016*, tiré de : https://www.lesaffaires.com/techno/technologie-de-l-information/les-quebecois-ont-depense-85-milliards--en-ligne-en-2016/594012/.

Lamb, C., Hair, J. F., McDaniel, C., Summers, J. et Gardiner, M. (2009). *Mktg*. Cengage Learning Australia, pp. 30-88.

Larson, P. D. et Halldorsson, A. (2004). Logistics versus supply chain management: An international survey. *International Journal of Logistics Research and Applications*, vol. 7, n° 1, pp. 17-31.

Lazar, M. (2019). *The ultimate Guide to Ecommerce Returns*, tiré de: https://www.readycloud.com/info/the-ultimate-guide-to-ecommerce-returns.

Lei, C., Zigang, Z. et Kaijun, L. (2009). Research on the Returns Policy Model of a Dual-Channel Supply Chain in E-commerce. *Paper presented at the International Symposium on Information Engineering and Electronic Commerce*, *IEEE*. Mai 16-17, Ternopil, Ukraine, pp. 21-25.

Leung, Y., Zhang, J. S. et Xu, Z. B. (2000). Clustering by scale-space filtering. *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, n° 12, pp. 1396-1410.

Li, G. et Li, W. (2015). The Analysis of Return Reverse Logistics Management Strategy Based on B2C Electronic Commerce. *Paper presented in 2015 International Conference on Economics, Social Science, Arts, Education and Management Engineering*. Atlantis Press. Octobre, Xi'an, Chine.

Liebert, W. et Schuster, H. G. (1989). Proper choice of the time delay for the analysis of chaotic time series. *Physics Letters A*, vol. 142, n° 2, pp. 107-111.

Lingam, Y. K. (2018). The role of Artificial Intelligence (AI) in making accurate stock decisions in E-commerce industry. *International Journal Advanced Research*. Ideas Innov. Technol, vol. 4, pp. 2281-2286.

Liu, D. (2014). Network site optimization of reverse logistics for E-commerce based on genetic algorithm. *Neural Computing and Applications*, vol. 25, n° 1, pp. 67-71.

Liu, H. et Motoda, H. (2012). *Feature selection for knowledge discovery and data mining*, vol. 454. Springer Science & Business Media, pp. 3-20.

Liu, J., Cheng, Q., Zheng, Z. et Qian, M. (2002). A DTW-based probability model for speaker feature analysis and data mining. *Pattern Recognition Letters*, vol. 23, n° 11, pp. 1271-1276.

Lou, Y., Ao, H. et Dong, Y. (2015). Improvement of Dynamic Time Warping (DTW) Algorithm. *Paper presented at the 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), IEEE*, Août 18-24, pp. 384-387.

Lu, Y. K. et Perron, P. (2010). Modeling and forecasting stock return volatility using a random level shift model. *Journal of Empirical Finance*, vol. 17, n° 1, pp. 138-156.

Luttwak, E. (1971), *A Dictionary of Modern War*, Harper & Row, New York, NY, pp. 680.

Ma, F. (2010). The Study on Reverse Logistics for E-Commerce. *Paper presented at the 2010 International Conference on Management and Service Science, IEEE*. Août 24-26, Wuhan, Chine, pp. 1-4.

Ma, J. et Kim, H. M. (2016). Predictive model selection for forecasting product returns. *Journal of Mechanical Design*, Transactions of the ASME, vol. 138, n° 5, 054501.

Maindonald, J. H. (2009). Time Series Analysis With Applications in R, Second Edition by Jonathan D. Cryer, Kung-Sik Chan. *International Statistical Review*, vol. 77, n° 2, pp. 300-301.

Marketingcharts (2015). *Global Retail and E-commerce Sales Forecast*, 213-2018, tiré de: https://www.marketingcharts.com/industries/retail-and-e-commerce-49733.

Makhabel, B. (2015). *Learning Data Mining with R*. Olton Birmingham, United Kingdom. Packt Publishing, Limited, pp. 73-95.

Malca, G. (2001). El comercio electrónico. *Repositorio de la Universidad del Pacifico*, vol. 1, Lima, Pérou.

McGee, M. et Yaffee, R. A. (2000). *An Introduction to Time Series Analysis and Forecasting: With Applications of SAS® and SPSS®*. Elsevier, pp. 23-44.

Micu, A., Micu, A. E., Geru, M. et Lixandroiu, R. C. (2017). Analyzing user sentiment in social media: Implications for online marketing strategy. *Psychology & Marketing*, vol. 34, n° 12, pp. 1094-1100.

Mitchell, T. M. (1997). Does machine learning really work?. *Artificial Intelligence magazine,* vol. 18, n° 3, pp. 11.

Mohapatra, S. (2013). E-commerce Strategy. *In E-Commerce Strategy*. Springer, Boston, MA. pp. 155-171

Monterroso, E. (2015). *El proceso logístico y la gestión de la cadena de abastecimiento*, tiré de: http://www.unlu.edu.ar/~ope20156/pdf/logistica.pdf

Mora, L.A. (2010*) KPI Los indicadores claves del desempeño logístico*, tiré de: http://www.fesc.edu.co/portal/archivos/e_libros/logistica/ind_logistica.pdf

Morel, M., Achard, C., Kulpa, R. et Dubuisson, S. (2018). Time-series averaging using constrained dynamic time warping with tolerance. *Pattern Recognition*, vol. 74, pp. 77-89.

Morrell, A. L. (2001). The forgotten child to the supply chain. *Modern Materials Handling*, vol. 56, n° 6, pp. 33-36.

Mu, Y., Liu, X., Yang, Z. et Liu, X. (2017). A parallel C4.5 decision tree algorithm based on MapReduce. *Concurrency and Computation: Practice and Experience*, vol. 29, n° 8, pp. 4015.

Murray, P. W., Agard, B. et Barajas, M. A. (2018). ASACT - Data preparation for forecasting: A method to substitute transaction data for unavailable product consumption data. *International Journal of Production Economics*, vol. 203, pp. 264-275.

O'connel, B. (2002). B2B.com-Ganhando dinheiro no E-Commerce Business-to-Business. *ISBN*, vol. 85, pp. 1244-1247.

Olorunniwo F.O. et Li X., (2010). Information sharing and collaboration practices in reverse logistics. *Supply Chain Management: An International Journal*, vol. 15, n° 6, pp. 454-462.

Orendorff, A. (2019). *Ecommerce Fashion Industry: Statistics, Trends & Strategy*, tiré de: https://www.shopify.com/enterprise/ecommerce-fashion-industry.

Pal, A. et Prakash, P. K. S. (2017). Practical Time Series Analysis: Master Time Series Data Processing, *Visualization, and Modeling using Python*. Packt Publishing Ltd, pp. 16-34.

Pandya, R. et Pandya, J. (2015). C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications*, vol. 117, pp. 18-21.

Pang, S.L. et Gong, J.Z. (2009). C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. *Systems Engineering - Theory & Practice*, vol. 29, n° 12, pp. 94-104.

Patil, N., Lathi, R. et Chitre, V. (2012). Comparison of C5. 0 & CART classification algorithms using pruning technique. *International Journal Engineering Research & Technology*, vol. 1, pp. 1-5.

Pearson, T. et Wegener, R. (2013). *Big data: the organizational challenge*. Bain and Company.

Pedrycz, W. et Chen, S. (2013). Time Series Analysis, Modeling and Applications. *A Computational Intelligence Perspective, pp. 31-76.*

Petropoulos, F., Kourentzes, N. et Nikolopoulos, K. (2016). Another look at estimators for intermittent demand. *International Journal of Production Economics*.

Pourakbar, M., van der Laan, E. et Dekker, R. (2014). End-of-Life Inventory Problem with Phaseout Returns. *Production and Operations Management*, vol. 23, n° 9, pp. 1561-1576.

Qi, M. et Zhang, G. P. (2008). Trend Time–Series Modeling and Forecasting With Neural Networks. *IEEE Transactions on Neural Networks*, vol. 19, n° 5, pp. 808-816.

Quinlan, J. R. (1986). *Induction of decision trees. Machine learning*, vol. 1, n° 1, pp. 81-106.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, vol. 28, n° 1, pp. 71-72.

Quinlan, J. R., (1993) *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, Inc., pp. 1-25.

Quinlan, J. R. et Rivest, R. L. (1989). Inferring decision trees using the minimum description lenght principle. *Information and computation*, vol. 80, n° 3, pp. 227-248.

Rabiner, L. R., Juang, B. H. et Lee, C. H. (1996). *An overview of automatic speech recognition. In Automatic Speech and Speaker Recognition*. Springer, Boston, MA. pp. 155-171.

Raman, V., et Hellerstein, J. (2001). Potter's Wheel: An Interactive Data Cleaning System, vol. 1, pp. 381-390.

Ramanathan, R. (2011). An empirical analysis on the influence of risk on relationships between handling of product returns and customer loyalty in E-commerce. *International Journal of Production Economics*, vol. 130, n° 2, pp. 255-261.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q. et Keogh, E. (2013). Addressing big data time series: Mining trillions of time series subsequences under dynamic time warping. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 7, n° 3, pp. 1-31.

Ratanamahatana, C. A. et Keogh, E. (2005). Three myths about dynamic time warping data mining. *In Proceedings of the 2005 SIAM International Conference on Data Mining.* Society for Industrial and Applied Mathematics, pp. 506-510.

Ramírez, A. C. (2009). *Manual de la gestión logística del transporte y distribución de mercancías.* Ediciones Uninorte, Barranquilla, Colombia, pp. 1-20.

Reig Torregrosa, R. (2017). *Percepción del e-commerce en el sector textil español.* Repositorio de Universidad de Alicante, Valencia, Espagne.

Rice, S. O. (1963). Noise in FM receivers. In Symposium *Proc. Time Series Analysis, 1963*, pp. 395-422.

Rogers, D. S. et Tibben-Lembke, R. (2001). An examination of reverse logistics practices. *Journal of business logistics*, vol. 22, n° 2, pp.129-148.

Rodriguez, A. et Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, vol. 344, n° 6191, pp.1492-1496.

Rouse, M. (2012). *E-commerce (electronic commerce)*, tiré de: https://searchcio.techtarget.com/definition/e-commerce.

Sakoe, H. et Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, n° 1, pp. 43-49.

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, vol. 16, n° 3, pp. 235-240.

Sato, H., Sagisaka, Y., Kogure, K. et Sagayama, S. (1982). Investigation in Japanese text-to-speech conversion. *Trans. of the Committee on Speech Research, S82*, vol. 8.

Seewald, A. K., Wernbacher, T., Pfeiffer, A., Denk, N., Platzer, M., Berger, M. et Winter, T. (2019). Towards Minimizing e-Commerce Returns for Clothing. Paper presented at the *ICAART International Conference on Agents and Artificial Intelligence*, vol. 2, pp. 801-808.

Sheikh, A.-S., Guigourès, R., Koriagin, E., Ho, Y. K., Shirvany, R., Vollgraf, R. et Bergmann, U. (2019). A deep learning system for predicting size and fit in fashion e-commerce. *Paper presented at the Proceedings of the 13th ACM Conference on Recommender Systems*, Copenhagen, Denmark, pp. 110-118.

Shivakumar, S. K. et Suresh, P. V. (2014) Maximizing Knowledge Management returns in e-commerce. *Paper presented at the 2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, Mars 5-7, New Delhi, Inde, pp. 545-550.

Sieranoja, S. et Fränti, P. (2019). Fast and general density peaks clustering. *Pattern Recognition Letters*, vol. 128, pp. 551-558.

Simpson, J., Ohri, L. et Lobaugh, K. (2016). The new digital divide: The future of digital influence in retail, tiré de: http://dupress.deloitte.com/dup-us-en/industry/retail-distribution/digital-divide-changing-consumer-behavior.html.

Sorensen, E. H., Miller, K. L. et Ooi, C. K. (2000). The decision tree approach to stock selection. *The Journal of Portfolio Management*, vol. 27, n° 1, pp. 42-52.

Teng, H.-M., Hsu, P.-H., Chiu, Y. et Wee, H. M. (2011). Optimal ordering decisions with returns and excess inventory. *Applied Mathematics and Computation*, vol. 217, n° 22, pp. 9009-9018.

Totonchi J. et Manshady K. (2012). Relationship between Globalization and E-Commerce. *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 2, n° 1, pp. 83.

Vallejo, D. et Tenalanda, G. (2012). Minería de datos aplicada en la detección de intrusos. *Ingenierías USBMed*, vol. 3, n° 1, pp, 50-61.

Velazquez, R. et Chankov, S. M. (2019). Environmental Impact of Last Mile Deliveries and Returns in Fashion E-Commerce: A Cross-Case Analysis of Six Retailers. *Paper presented at the 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. Décembre 15-18, Macao, Chine, pp. 1099-1103.

Verma, M., Srivastava, M., Chack, N., Diswar, A. K. et Gupta, N. (2012). A comparative study of various clustering algorithms in data mining. *International Journal of Engineering Research and Applications (IJERA)*, vol. 2, n° 3, pp. 1379-1384.

Van der Vlist, R., Taal, C. et Heusdens, R. (2019). Tracking Recurring Patterns in Time Series Using Dynamic Time Warping. *Paper presented at the 2019 27th European Signal Processing Conference (EUSIPCO)*. Septembre 2-6, Coruna, Espagne, pp. 1-5.

Wang, J. L. et Chan, S. H. (2006). Stock market trading rule discovery using two-layer bias decision tree. *Expert Systems with Applications*, vol. 30, n° 4, pp. 605-611.

Wang, K. et Gasser, T. (1997). Alignment of curves by dynamic time warping. *Ann. Statist.*, vol. 25, n° 3, pp. 1251-1276.

Wang, W. (2015). A Decision Method for Returns Logistics Based on the Customer's Behaviour in E-commerce. *Procedia Computer Science*, vol. 60, pp. 1506-1515.

Wang, X., Xie, J. et Fan, Z.-P. (2020). B2C cross-border E-commerce logistics mode selection considering product returns. *International Journal of Production Research*, pp. 1-20.

Wang, Y., Lei, P., Zhou, H., Wang, X., Ma, M. et Chen, X. (2014). Using DTW to measure trajectory distance in grid space. *In 2014 4th IEEE International Conference on Information Science and Technology*, Avril, pp. 152-155.

Ware, V. S. et Bharathi, H. N. (2013). Study of density based algorithms. *International Journal of Computer Applications*, vol. 69, n° 26, pp. 1-4.

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition*, vol. 38, n° 11, pp. 1857-1874.

Wei, C., Li, Y. et Cai, X. (2011). Robust optimal policies of production and inventory with uncertain returns and demand. *International Journal of Production Economics*, vol. 134, n° 2, pp. 357-367.

Weiss, G. M., McCarthy, K. et Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. *Dmin*, vol. 7, n° 35-41, pp. 24.

Witten, I. H., Frank, E., Hall, M. A. et Pal, C. J. (2017). Chapter 2 - Input: Concepts, instances, attributes. *Data Mining (Fourth Edition*. Morgan Kaufmann, pp. 43-66.

XiaoYan, Q., Han, Y., Qinli, D. et Stokes, P. (2012). Reverse logistics network design model based on e-commerce. *International Journal of Organizational Analysis*, vol. 20, n° 2, pp. 251-261.

Yadollahi, E., Aghezzaf, E.-H. et Raa, B. (2017). Managing inventory and service levels in a safety stock-based inventory routing system with stochastic retailer demands. *Applied Stochastic Models in Business and Industry*, vol. 33, n° 5, pp. 555-555.

Yang, H., Xu, B., Zhou, Y., Zhang, J. et Biao, D. (2010). Return in B2C E-Commerce Enterprises. Paper presented at the 2010 *Third International Conference on Business Intelligence and Financial Engineering*. Août 13-15, pp. 95-98.

Yu, D., Yu, X., Hu, Q., Liu, J. et wu, A. (2011). Dynamic time warping constraint learning for large margin nearest neighbor classification. *Information Sciences*, vol. 181, pp. 2787-2796.

Yuan, M., Buttenfield, B., Gahegan, M. et Miller, H. (2004). Geospatial data mining and knowledge discovery. *A research agenda for geographic information science*, vol.3, pp. 365.

Yuan, X. M., Tan, Z. L. et Bhullar, A. S. (2014). Optimal inventory policies for remanufacturing inventory systems with multiple returns. Paper presented at the 2014 *IEEE International Conference on Industrial Engineering and Engineering Management*. Décembre 9-12, pp. 1382-1388.

Zermati, P. et Mocellin, F. (2005). *Pratique de la gestion des Stocks*. Dunod, Paris, France.

Zhao, J. et Itti, L. (2018). shapeDTW: Shape Dynamic Time Warping. *Pattern Recognition*, vol. 74, pp. 171-184.

Zhu, Y., Li, J., He, J., Quanz, B. et Deshpande, A. (2018). A Local Algorithm for Product Return Prediction in E-Commerce. *In International Joint Conference of Artificial Intelligence*, pp. 3718-3724.

Zucchini, W., MacDonald, I. L. et Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*, Second Edition. Oakville, United States: CRC Press LLC, pp. 331-338.