REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ 20 AOÛT 1955 -SKIKDA

Faculté des Sciences

Département d'Informatique

Mémoire de fin d'études en vue de l'obtention du diplôme

De Master en Informatique

Spécialité : Systèmes d'Information et Applications avancées (SIAA)

Thème

Segmentation client par machine learning : application marketing

Réalisé par :

- ✓ BOUGHIOUT Youcef
- ✓ TOURECHE Oussama

Encadré par :

M^{me} MAGROUN Hanane

Session : *Juin 2025*

Remerciements Et Dédicaces

Nous tenons à exprimer toute nos reconnaissances a notre directrice de mémoire, madame **Magroun Hanane**. *Nous la remercions de nous avoir encadrés, orienté, aidé et conseillé. Nous remercions également les membres de jury qui ont accepté à évaluer notre travail.*

Je remercie nos chers parents qui ont toujours été la pour nous, et pour leurs encouragements.

Enfin, nous remercions nos amis qui ont toujours été la pour nous, leur encouragements ont été d'une grande aide.

Abstract

Companies, particularly those in the e-commerce and retail sectors, collect extensive demographic and financial data about their customers that provide crucial insights into consumer behavior patterns and purchasing power. This thesis addresses the problem of customer segmentation using the K-means algorithm to identify and analyze different customer categories based on demographic and economic characteristics in a targeted marketing strategy. The study problem is the difficulty for companies to effectively categorize their diverse customer base and tailor their marketing approaches without proper segmentation based on key demographic and financial indicators. The proposed solution consists of applying the K-means algorithm on customer data including age, gender, annual income, and spending score to create meaningful customer clusters. The obtained results reveal distinct customer segments with spending profiles, enabling companies to develop targeted marketing strategies and optimize product offerings for each identified customer group.

Keywords: Machine Learning, Customer Segmentation, K-means, Marketing, Demographics, Spending Behavior.

ملخص:

تقوم الشركات، خاصة تلك العاملة في قطاعي التجارة الإلكترونية والبيع بالتجزئة، بجمع بيانات ديموغرافية ومالية واسعة حول عملائها والتي توفر رؤى حاسمة حول أنماط سلوك المستهلكين وقدرتهم الشرائية. تتناول هذه المذكرة مشكلة تجميع العملاء باستخدام خوارزمية K-means لتحديد وتحليل فئات مختلفة من العملاء بناءً على الخصائص الديموغرافية والاقتصادية في إستراتيجية تسويق مستهدفة. مشكلة الدراسة هي صعوبة الشركات في تصنيف قاعدة عملائها المتنوعة بفعالية وتكييف مناهجها التسويقية دون تجميع مناسب يعتمد على المؤشرات الديموغرافية والمالية الرئيسية. يتمثل الحل المقترح في تطبيق خوارزمية K-means على بيانات العملاء متضمنة العمر والجنس والدخل السنوي ونتيجة الإنفاق لإنشاء مجموعات عملاء ذات معنى. تكشف النتائج المحصل عليها عن قطاعات عملاء متميزة ذات إنفاق محدد، مما يمكن الشركات من تطوير استراتيجيات تسويقية مستهدفة وتحسين عروض المنتجات لكل مجموعة عملاء محددة.

الكلمات المفتاحية: تعلم الآلة، تجميع العملاء، K-means، التسويق، الديموغرافيا، سلوك الإنفاق.

Résumé :

Les entreprises, en particulier celles des secteurs du commerce électronique et de la vente au détail, collectent des données démographiques et financières étendues sur leurs clients qui fournissent des insights cruciaux sur les modèles de comportement des consommateurs et leur pouvoir d'achat. Ce mémoire aborde le problème de la segmentation des clients en utilisant l'algorithme K-means pour identifier et analyser différentes catégories de clients basées sur les caractéristiques démographiques et économiques dans une stratégie marketing ciblée. Le problème d'étude est la difficulté pour les entreprises à catégoriser efficacement leur base de clientèle diversifiée et à adapter leurs approches marketing sans une segmentation appropriée basée sur des indicateurs démographiques et financiers clés. La solution proposée consiste à appliquer l'algorithme K-means sur des données clients incluant l'âge, le sexe, le revenu annuel et le score de dépenses pour créer des clusters clients significatifs. Les résultats obtenus révèlent des segments clients distincts avec des profils de dépenses spécifiques, permettant aux entreprises de développer des stratégies marketing ciblées et d'optimiser les offres produits pour chaque groupe de clients identifié.

Mots clés : Apprentissage automatique, Segmentation clients, K-means, Marketing, Démographie, Comportement de dépenses.

Sommaire

Liste des figures

Liste des tableaux

Introduction générale

Introduction générale

Dans un environnement économique de plus en plus concurrentiel, la compréhension approfondie de la clientèle constitue un enjeu stratégique majeur pour les entreprises modernes. L'évolution rapide des technologies de l'information et l'explosion du volume de données disponibles ont transformé la façon dont les organisations appréhendent leurs relations avec leurs clients. Les secteurs du commerce électronique et de la vente au détail, en particulier, génèrent quotidiennement d'importants volumes de données démographiques et transactionnelles qui recèlent un potentiel considérable pour l'optimisation des stratégies commerciales.

Cette révolution numérique a permis aux entreprises de collecter et de stocker des informations détaillées sur les habitudes de consommation, les préférences et les caractéristiques socio-économiques de leur clientèle. Toutefois, l'exploitation efficace de ces données massives demeure un défi complexe qui nécessite l'adoption d'approches analytiques sophistiquées. La capacité à transformer ces données brutes en insights actionnables représente désormais un avantage concurrentiel déterminant dans la réussite commerciale.

Malgré la richesse des données clients disponibles, de nombreuses entreprises peinent à exploiter pleinement ce potentiel informationnel. La principale difficulté réside dans la capacité à catégoriser efficacement une base de clientèle souvent hétérogène et complexe. Sans une segmentation appropriée basée sur des indicateurs démographiques et financiers pertinents, les entreprises se trouvent dans l'incapacité de personnaliser leurs approches marketing et d'adapter leurs offres aux besoins spécifiques de chaque groupe de consommateurs.

Cette problématique de segmentation inadéquate engendre plusieurs conséquences néfastes : dispersion des efforts marketing, allocation inefficiente des ressources publicitaires, inadéquation entre l'offre produit et les attentes clients, et ultimement, une perte de compétitivité face aux entreprises ayant su développer des stratégies de ciblage plus précises. La question centrale qui se pose est donc : comment les entreprises peuvent-elles identifier et analyser de manière systématique les différents segments de leur clientèle pour optimiser leurs stratégies marketing ?

L'objectif principal de ce mémoire consiste à développer et valider une méthodologie de segmentation clientèle basée sur l'algorithme d'apprentissage automatique K-means, permettant aux entreprises d'identifier des groupes homogènes de clients partageant des caractéristiques démographiques et comportementales similaires.

Les objectifs spécifiques de cette recherche s'articulent autour de plusieurs axes :

- ✓ Analyser les variables démographiques et financières pertinentes pour la segmentation clientèle, notamment l'âge, le sexe, le revenu annuel et le score de dépenses

- ✓ Appliquer l'algorithme K-means sur un jeu de données clients représentatif pour identifier des clusters significatifs
- ✓ Évaluer la qualité et la pertinence des segments obtenus en termes de cohérence interne et de différenciation externe
- ✓ Interpréter les profils de dépenses spécifiques à chaque segment identifié
- ✓ Formuler des recommandations stratégiques pour l'adaptation des approches marketing ciblées

Cette étude adopte une approche quantitative basée sur l'analyse de données et l'application d'algorithmes d'apprentissage automatique non supervisé. La méthodologie s'appuie sur l'utilisation de l'algorithme K-means, une technique de clustering particulièrement adaptée à la segmentation clientèle en raison de sa capacité à partitionner efficacement des données multidimensionnelles en groupes homogènes.

Le processus méthodologique comprend plusieurs étapes clés : la collecte et la préparation des données clients, l'analyse exploratoire des variables démographiques et financières, l'application de l'algorithme K-means avec optimisation du nombre de clusters, l'évaluation de la qualité des segments obtenus, et enfin l'interprétation business des résultats pour formuler des recommandations actionables.

Ce mémoire s'organise en plusieurs chapitres complémentaires qui permettent d'aborder progressivement la problématique de segmentation clientèle. Après cette introduction, une revue de littérature présentera les fondements théoriques de la segmentation marketing et les applications de l'apprentissage automatique dans ce domaine. Le chapitre méthodologique détaillera l'approche analytique adoptée et les outils utilisés. Les résultats obtenus feront l'objet d'une présentation et d'une analyse approfondie, suivies d'une discussion sur leurs implications managériales. Enfin, une conclusion synthétisera les contributions de cette recherche et ouvrira sur des perspectives d'approfondissement futur.

Pour mener ce travail, nous avons organisé le document en trois chapitres :

- Le premier chapitre qui porte sur la segmentation client
- Le deuxième chapitre est dédié aux fondements de machine learning
- Le dernier chapitre est consacre l’étude de cas segmentation client par clustering K-Means

Chapitre 1
La segmentation client

1. Introduction

La segmentation client est un pilier du CRM moderne, permettant d'adapter l'offre commerciale aux besoins spécifiques de chaque profil. En regroupant les clients selon leurs comportements et caractéristiques, les entreprises optimisent leurs stratégies marketing. Ce chapitre explore les méthodes et enjeux de cette approche centrée sur la donnée.

2. CRM - Customer Relationship Management

Le Customer Relationship Management (CRM) se traduit par gestion de la relation client est une approche stratégique globale visant à établir, maintenir et renforcer des relations durables avec des clients sélectionnés. Il s'agit d'un processus intégratif qui unit le marketing, les ventes, le service client et la gestion de la chaîne logistique, afin d'optimiser la création de valeur tant pour l'entreprise que pour le client. Le CRM repose sur trois piliers fondamentaux : La technologie, qui permet la collecte, le stockage, l'analyse et l'exploitation des données client ; Les individus, c'est-à-dire les clients comme les employés, où l'implication des équipes internes est cruciale pour assurer l'efficacité de la stratégie ; Les processus métiers, qui doivent être réorientés vers une approche centrée client afin de personnaliser les interactions et anticiper les attentes. Ainsi, le CRM devient bien plus qu'un simple outil logiciel : il s'impose comme un levier stratégique essentiel pour créer des avantages concurrentiels durables [1]**.**

2.1. Les composantes du CRM

Le CRM ne se limite pas à un simple outil technologique. Il repose sur plusieurs composantes clés qui interagissent pour permettre une gestion efficace de la relation client. Ces composantes sont généralement regroupées comme suit :

- ✓ Le CRM analytique : il consiste à collecter et analyser les données relatives aux clients afin d'en extraire des informations pertinentes pour la prise de décision. Il repose sur des technologies de Data Mining, de tableaux de bord, et d'indicateurs de performance (KPI) **[1]**
- ✓ Le CRM opérationnel : il regroupe l'ensemble des processus automatisés liés aux fonctions commerciales, marketing et service client. Il comprend la gestion des campagnes, des ventes, des contacts et des réclamations, en veillant à la cohérence des interactions avec le client [1]

- ✓ Le CRM collaboratif : cette composante vise à favoriser l'échange d'informations entre les différents départements internes de l'entreprise ainsi qu'avec les partenaires externes (distributeurs, prestataires de services). L'objectif est d'assurer une vision unifiée du client à tous les niveaux de l'organisation. En intégrant ces trois dimensions, le CRM permet une approche globale, coordonnée et personnalisée de la gestion client [1]

2.2. Les étapes du processus CRM

Le processus CRM s'articule autour de quatre étapes fondamentales qui structurent l'ensemble des interactions avec le client tout au long de son cycle de vie. Ces étapes visent à créer une relation durable, rentable et mutuellement bénéfique : [1]

2.1.1. L'identification

Il s'agit de repérer et de collecter les données sur les clients actuels et potentiels. Cette phase permet de construire une base de données riche, essentielle pour toute action personnalisée.

2.1.2. L'attraction

Elle vise à capter l'attention des prospects à travers des campagnes ciblées, des offres personnalisées ou des stratégies multicanales. L'objectif est de convertir ces prospects en clients actifs.

2.1.3. La fidélisation

Une fois le client acquis, il devient crucial de maintenir la relation en renforçant la satisfaction, la confiance et l'attachement à la marque. Les programmes de fidélité, le service après-vente et la qualité des interactions jouent ici un rôle central.

2.1.4. Le développement

Cette dernière étape vise à accroître la valeur client sur le long terme en favorisant les ventes croisées (cross-selling), les ventes additionnelles (up-selling) et la recommandation active par le client. Ces quatre étapes, interconnectées, forment un cadre structuré qui guide l'entreprise dans la construction d'une stratégie relationnelle efficace et orientée vers la performance.

2.3. Les avantage du CRM

La mise en œuvre d'un système CRM efficace procure de nombreux avantages tant pour l'entreprise que pour le client. Ces bénéfices se manifestent à différents niveaux [1]:

2.3.1. Amélioration de la connaissance client

Le CRM centralise les données et permet une analyse fine des comportements, besoins et préférences, facilitant la prise de décision stratégique.

2.3.2. Personnalisation de l'offre

En segmentant la clientèle et en adaptant les propositions commerciales, l'entreprise renforce la pertinence de ses actions marketing et la satisfaction des clients.

2.3.3. Optimisation de la fidélisation

Grâce à un meilleur suivi de la relation client, l'entreprise peut anticiper les attentes, prévenir les départs, et accroître la durée de vie client (CLV). Augmentation de la rentabilité : en ciblant mieux les campagnes, en réduisant les coûts de gestion et en maximisant la valeur client, le CRM contribue à améliorer la performance économique globale.

2.3.4. Renforcement de la coordination interne

Les différentes équipes (vente, marketing, service client) accèdent à une vision unifiée du client, favorisant la cohérence des actions et la fluidité des échanges. Ainsi, le CRM s'impose comme un outil stratégique incontournable pour développer une relation client solide, rentable et durable.

2.3.5. Les outils CRM utilisés

L'évolution des technologies numériques a favorisé l'émergence de nombreux outils CRM qui permettent aux entreprises de gérer efficacement leurs relations clients. Ces solutions, qu'elles soient intégrées ou modulaires, offrent des fonctionnalités variées adaptées aux besoins spécifiques de chaque organisation. Parmi les outils les plus répandus, on retrouve [1] :

2.3.6. Sales force

Leader du marché mondial, il propose une plateforme complète de gestion de la relation client, incluant automatisation des ventes, service client et analytique avancé.

2.3.7. Microsoft Dynamics

Il combine CRM et ERP dans une même solution, facilitant l'intégration des processus métiers et des données clients.

2.3.8. HubSpot CRM

Particulièrement prisé par les PME, il offre une interface intuitive et des fonctionnalités gratuites pour la gestion des contacts, des leads et des campagnes marketing. Zoho CRM, SAP Customer Experience et Pipedrive sont également des solutions puissantes permettant de personnaliser les interactions, de suivre les performances commerciales, et d'améliorer la collaboration inter-équipes.

3. La segmentation client

La segmentation client désigne une démarche stratégique consistant à diviser la clientèle d'une entreprise en groupes homogènes selon des critères partagés tels que les comportements d'achat, les habitudes de consommation, les préférences ou encore la valeur économique. Cette classification permet de mieux comprendre les attentes spécifiques de chaque segment afin d'optimiser les actions marketing, commerciales et relationnelles. Introduite dans le domaine du marketing dès les années 1950 par Wendell R. Smith, la segmentation s'est imposée comme une pratique essentielle dans un environnement où la personnalisation est devenue un facteur de différenciation majeur. Elle permet aux entreprises d'affiner leur connaissance client, de cibler précisément leurs offres et de maximiser l'efficacité de leurs campagnes [2].

3.1. Les types de segmentation client

3.1.1. Segmentation basée sur la valeur

La segmentation basée sur la valeur consiste à classer les clients selon l'importance économique qu'ils représentent pour l'entreprise. Cette importance peut être mesurée à travers divers indicateurs tels que le chiffre d'affaires généré, la marge dégagée, ou encore la rentabilité sur le long terme. En identifiant les clients à forte valeur, l'entreprise peut prioriser ses efforts commerciaux et marketing, adapter son offre, et maximiser sa rentabilité. Cette approche est particulièrement pertinente dans des secteurs à forte intensité concurrentielle, où l'optimisation des ressources devient essentielle [3].

3.1.2. Segmentation comportementale

La segmentation comportementale repose sur l'observation et l'analyse des actions des clients. Elle tient compte de la fréquence d'achat, des catégories de produits préférées, du moment d'achat, du canal utilisé (en ligne ou en magasin), ou encore de la réactivité aux campagnes marketing. Cette méthode permet une compréhension fine des habitudes de consommation et offre la possibilité de personnaliser les interactions. Grâce aux données recueillies via les systèmes CRM ou les historiques de transaction, cette forme de segmentation est largement utilisée dans la vente au détail et le e-commerce [3].

3.1.3. Segmentation par propension

La segmentation par propension s'appuie sur des modèles prédictifs pour estimer la probabilité qu'un client réalise une action future : acheter un nouveau produit, répondre à une campagne, se désabonner, etc. Ces scores de propension sont générés grâce à des algorithmes de machine learning ou des méthodes statistiques classiques. Cette approche permet d'anticiper les comportements et de concevoir des actions marketing proactives et ciblées, réduisant ainsi le risque d'attrition et améliorant les taux de conversion [3].

3.1.4. Segmentation par fidélité

La segmentation par fidélité distingue les clients en fonction de leur niveau d'engagement envers l'entreprise ou la marque. Elle repose sur des critères tels que l'ancienneté, la fréquence de réachat, ou la participation à des programmes de fidélisation. Elle permet d'identifier des segments comme les clients « fidèles », « irréguliers », ou « inactifs ». Cette méthode est précieuse pour adapter les stratégies de rétention, stimuler l'engagement et prolonger la durée de vie client (Customer Lifetime Value) [3].

3.1.5. Segmentation socio-démographique et cycle de vie

Cette segmentation repose sur des données telles que l'âge, le sexe, le revenu, la situation familiale ou encore la profession. Elle permet de regrouper les clients selon des caractéristiques objectives, souvent faciles à collecter. Elle est utile pour adapter les produits et services à différentes étapes du cycle de vie, comme les jeunes actifs, les familles, ou les retraités. Bien qu'élémentaire, cette approche reste complémentaire à d'autres formes plus complexes de segmentation [3].

3.1.6. Segmentation par besoins et attitudes

Cette forme de segmentation cherche à regrouper les clients selon leurs attentes profondes, leurs motivations, leurs préférences ou encore leurs valeurs. Elle est généralement issue d'enquêtes qualitatives, d'études de marché ou d'analyses psychographiques. Elle offre un haut niveau de personnalisation et s'avère précieuse dans le développement de nouveaux produits ou pour affiner le positionnement d'une marque. Elle permet aussi de détecter des segments latents non visibles à travers les seules données transactionnelles [3].

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

L'importance de la segmentation client dans la stratégie marketing Dans un environnement économique de plus en plus concurrentiel, la segmentation client s'impose comme un levier stratégique essentiel pour toute organisation orientée vers le client. Elle permet d'identifier les groupes de clients ayant des besoins, des attentes ou des comportements similaires, afin de proposer des offres ciblées et pertinentes. En adaptant les produits, les services et les campagnes de communication à chaque segment identifié, les entreprises peuvent améliorer significativement leur efficacité marketing, renforcer la fidélisation et maximiser le retour sur investissement (ROI). De plus, une segmentation bien exécutée facilite la prise de décision, oriente le développement commercial et soutient la différenciation sur le marché. Elle devient ainsi un outil fondamental dans la construction d'une relation client durable et rentable, en particulier dans les secteurs où l'expérience personnalisée constitue un facteur de compétitivité clé [3].

3.2.1. Le secteur bancaire

Le secteur bancaire, confronté à une transformation numérique rapide et à une concurrence accrue, adopte la segmentation client comme levier d'optimisation stratégique. Cette pratique permet aux établissements financiers de mieux comprendre les profils d'emprunteurs et d'adapter les offres en conséquence. En catégorisant les clients selon des critères tels que la valeur économique, le comportement de paiement ou la fidélité, les banques peuvent personnaliser leurs services et améliorer la rentabilité tout en maîtrisant les risques. La segmentation devient ainsi un outil clé dans la gestion de la relation client bancaire, facilitant aussi bien les actions marketing ciblées que les stratégies de rétention à long terme [4].

3.2.2. Le secteur de détail

Dans le domaine du commerce de détail, la segmentation client représente un levier incontournable pour renforcer l'efficacité des programmes de fidélité et personnaliser l'expérience d'achat. Les entreprises de détail s'appuient sur l'analyse des comportements d'achat pour regrouper les clients selon leur fréquence de visite, leurs dépenses, ou encore leur sensibilité aux promotions. Cette démarche permet d'adapter les offres commerciales aux attentes spécifiques de chaque segment, de cibler les actions marketing avec plus de précision, et d'optimiser la gestion des ressources. En segmentant intelligemment leur clientèle, les enseignes peuvent ainsi améliorer la fidélisation, accroître la satisfaction, et renforcer leur position concurrentielle sur le marché [4].

3.2.3. Le secteur e-commerce

Le secteur du e-commerce, fortement digitalisé et riche en données, offre un terrain idéal pour la mise en œuvre de stratégies de segmentation avancées. Grâce à l'analyse des comportements en ligne et à l'historique des transactions, les plateformes peuvent identifier des profils variés de clients, allant des acheteurs occasionnels aux ambassadeurs de marque. La segmentation dans le e-commerce ne se limite pas aux aspects transactionnels. Elle intègre également des dimensions sociales telles que la capacité d'un client à recommander le service, à générer du trafic, ou à influencer d'autres consommateurs via des programmes de parrainage ou des réseaux sociaux. Cette double lecture – économique et sociale – permet de mettre en place des actions marketing beaucoup plus précises et personnalisées [4].

4. Conclusion

La segmentation client, lorsqu'elle est bien exploitée, devient un levier puissant pour personnaliser le parcours client et renforcer la fidélisation. Ces principes trouvent leur prolongement naturel dans l'ère du machine learning, où les algorithmes comme le K-means automatisent et affinent cette segmentation. C'est ce que nous verrons dans le prochain chapitre.

Chapitre 2
Fondements du Machine Learning

1. Introduction

Le machine learning repose sur des algorithmes capables d'apprendre à partir de données sans programmation explicite. Ses fondements combinent statistiques, optimisation et informatique pour résoudre des problèmes complexes. Ce chapitre explore les principes clés qui sous-tendent cette discipline révolutionnaire

2. l'Intelligence Artificielle

L'intelligence artificielle (IA) est aujourd'hui au cœur d'une transformation majeure qui touche presque tous les secteurs d'activité, de la santé au transport, en passant par l'industrie, la finance, et les services numériques. L'IA se définit comme l'ensemble des théories et des techniques développant des programmes informatiques capables de simuler certains traits de l'intelligence humaine, tels que l'apprentissage, le raisonnement ou la perception

Historiquement, l'IA a émergé dès les années 1950 avec les travaux pionniers d'Alan Turing et la création des premiers neurones artificiels comme le perceptron de Frank Rosenblatt . Depuis, cette discipline a connu plusieurs révolutions, en particulier grâce à l'essor de la puissance de calcul, de l'accès à de grandes quantités de données, et à l'émergence de techniques d'apprentissage automatique (Machine Learning) et d'apprentissage profond (Deep Learning).

Parmi ses composantes majeures, l'apprentissage automatique permet aux machines d'améliorer leur performance à travers l'expérience, sans programmation explicite. Il se décline en plusieurs formes, dont l'apprentissage supervisé, non supervisé et par renforcement . Ces techniques se concrétisent dans des applications variées, comme la détection d'obstacles pour les systèmes embarqués , la détection d'objets et de pièces détachées dans l'industrie automobile , ou encore le diagnostic médical assisté par IA dans les ambulances intelligentes [5] [6] [7] [8].

2.1. Historique

En **(1943-1956)** Warren McCulloch et Walter Pitts ont publié le premier document sur l'intelligence artificielle en 1943, proposant alors un neurone artificiel. Ensuite, la machine de Turing, conçue par Alan Turing en 1950, a servi dans les tests. L'aptitude de l'appareil à interagir avec les êtres humains.

Par la suite, il a publié en 1951 un article intitulé « Computing Machinery and Intelligence » où il a suggéré une simulation qui serait plus tard connue sous le nom de Test de Turing. Allen Newell et Herbert A. Simon ont conçu le tout premier programme d'intelligence artificielle en 1955. John

McCarthy, un informaticien américain, a conçu l'intelligence artificielle lors de la Conférence de Dartmouth en 1956. Pour le premier temps, l'IA a été reconnue comme un domaine d'étude.

En **(1966-1996)** ELIZA, le premier chatbot, a été inventé par Joseph Weizenbaum en 1966. En 1972, le Japon a mis au point WABOT-1, qui est le premier robot humain intelligent au monde . En 1996, IBM Deep Blue, couronné champion du monde d'échecs, a battu Gary Kasparov, devenant ainsi le premier ordinateur capable de triompher d'un champion d'échecs.

En (**2006-2022**) La première intelligence artificielle à faire son apparition dans le domaine commercial remonte à 2006. Des entreprises telles que Facebook, Twitter et Netflix utilisent désormais l'intelligence artificielle. En 2018, Google a lancé « Duplex », une application d'intelligence artificielle qui fait office d'assistant virtuel. Depuis sa première mise en circulation publique le 30 novembre 2022, ChatGPT La remarquable aptitude de ChatGPT à réaliser des missions sophistiquées dans le secteur éducatif, comme cette progression en intelligence artificielle, paraît transformer radicalement les méthodes éducatives actuelles [9] [10] .

2.2. Définition d'intelligence artificielle

Même si l'intelligence artificielle est perçue comme le dernier domaine novateur en matière de progrès technologique, le terme a été forgé en août 1956 par John McCarthy. Il a alors défini l'IA comme la science et l'ingénierie de la création de machines intelligentes. Actuellement, l'IA est étudiée et mise en œuvre dans une multitude de secteurs couvrant toutes sortes de tâches cognitives, ce qui en fait un champ d'étude véritablement universel. On a souvent affirmé que l'intelligence artificielle pourrait supplanter les êtres humains. Une autre faction de théoriciens estime qu'elle amplifie les facultés humaines et nous autorise à accomplir notre tâche de façon efficiente [9].

2.3. Approches de l'AI

On peut aborder l'IA à travers divers paradigmes et techniques, chacun présentant ses propres avantages et inconvénients. Voici quelques exemples illustratifs :

2.3.1. Machine Learning (ML)

L'apprentissage automatique (ML) se base sur des algorithmes capables d'extraire des savoirs à partir de vastes volumes de données et de perfectionner progressivement leurs performances au fur et à mesure qu'ils accumulent davantage d'informations. Ces méthodes englobent des stratégies comme l'apprentissage supervisé, non supervisé, ainsi que

l'apprentissage par renforcement. Par exemple, dans le domaine financier, l'intelligence artificielle est employée pour identifier les fraudes et anticiper les variations du marché. Dans le domaine médical, les modèles d'apprentissage servent à anticiper des diagnostics en se basant sur des informations cliniques [11].

2.3.2. Réseaux de neurones artificiels (RNA)

Les systèmes de réseaux de neurones artificiels (RNA), qui imitent le fonctionnement du cerveau humain, sont des structures informatiques sophistiquées capables d'apprendre et de s'adapter à partir de données. L'évolution des architectures de réseaux de neurones a conduit du perceptron simple aux réseaux profonds (Deep Neural Networks, DNN), incorporant plusieurs couches cachées. Ces structures se révèlent particulièrement performantes pour des missions comme l'identification d'images, la traduction automatique et la production de voix.

Les progrès récents dans le domaine des réseaux neuronaux convolutifs (CNN) et des réseaux neuronaux récurrents (RNN) ont grandement optimisé les performances des systèmes en matière de vision par ordinateur et de traitement du langage naturel (NLP) [6].

2.3.3. Logique symbolique

Cette méthode consiste à utiliser des symboles et des règles logiques pour exprimer des connaissances et effectuer des déductions. Les systèmes experts, qui visent à reproduire la réflexion humaine dans des secteurs tels que la médecine ou l'ingénierie, font souvent appel à la logique symbolique. Les systèmes qui se fondent sur la logique symbolique fournissent une interprétation formelle des données, favorisant ainsi la transparence et le suivi des décisions. Toutefois, leur contrainte majeure est la gestion des situations où les normes ne sont pas explicitement établies ou en présence d'incertitudes [6].

2.3.4. Systèmes basés sur des règles

Les systèmes régis par des règles opèrent en se basant sur un ensemble de directives conditionnelles qui provoquent des actions déterminées dès lors que des conditions spécifiques sont satisfaites. On utilise couramment ces systèmes dans des applications d'automatisation industrielle, de gestion des stocks ou de contrôle automatique, où il est primordial d'avoir une prise de décision précise et rapide. Même si ces systèmes sont performants pour des activités répétitives, ils font défaut de souplesse lorsqu'ils sont confrontés à des situations imprévues ou compliquées [12].

3. Intelligence computationnelle

L'intelligence computationnelle englobe plusieurs techniques bio-inspirées telles que les algorithmes génétiques, les systèmes flous et les réseaux de neurones. Ces techniques sont particulièrement efficaces pour traiter des problématiques d'optimisation complexes et dénicher des solutions imitant les processus naturels. Par exemple, les algorithmes génétiques imitent le mécanisme de sélection naturelle en parcourant un large éventail de solutions possibles, ce qui les rend appropriés pour des problématiques d'optimisation dans des secteurs tels que la logistique ou le design de produits [12].

2.4. Applications de l'AI

L'intelligence artificielle (IA) a fait son entrée dans plusieurs secteurs, fournissant des solutions novatrices et révolutionnant les méthodes en place dans une variété de domaines. Voici quelques illustrations de l'utilisation de l'IA :

2.4.1. Santé

L'intelligence artificielle a transformé le domaine de la santé en facilitant le diagnostic médical par ordinateur, l'élaboration de médicaments et l'examen d'images médicales. Par exemple, des systèmes d'intelligence artificielle tels qu'IBM Watson16 examinent des dossiers médicaux complexes afin de suggérer des thérapies sur mesure. On utilise aussi des algorithmes de reconnaissance d'image pour identifier les irrégularités dans les radiographies ou les IRM, augmentant ainsi l'exactitude des diagnostics, spécifiquement pour le dépistage précoce de maladies telles que le cancer [13].

2.4.2. Robotique

L'intelligence artificielle est essentielle pour l'élaboration de robots autonomes aptes à interagir avec leur environnement. Des robots tels que Spot de Boston Dynamics sont déployés pour la gestion d'objets complexes et l'examen des infrastructures dans des contextes à risque [14].

2.4.3. Transport

L'intelligence artificielle a révolutionné le secteur des transports grâce à l'essor des véhicules autonomes et à l'amélioration de la gestion du trafic. Des sociétés telles que Tesla et Waymo élaborent des voitures autonomes qui se servent de réseaux neuronaux pour analyser les

informations provenant des capteurs et effectuer des choix en direct. En outre, des dispositifs de contrôle de la circulation alimentés par l'IA modifient les signaux lumineux selon l'état du trafic, atténuant ainsi les congestions dans les smart cities [15].

2.4.4. Industrie

Dans les contextes industriels, l'IA est mise en œuvre pour la surveillance des processus, la maintenance anticipative ainsi que l'optimisation de la production. Par exemple, General Electric a recours à des systèmes d'intelligence artificielle pour superviser ses machines et anticiper les pannes avant leur occurrence, ce qui permet de diminuer les périodes d'inactivité et les dépenses d'entretien. En outre, les modèles de prévision améliorent les niveaux d'inventaire en se basant sur les tendances de la demande [15].

2.4.5. Éducation

L'intelligence artificielle facilite la mise en place de systèmes de tutorat intelligent et la personnalisation de l'apprentissage. Des services comme Khan Academy se servent d'algorithmes d'intelligence artificielle pour personnaliser le contenu éducatif en fonction des résultats des élèves. Les instruments d'analyse prédictive contribuent aussi à repérer les étudiants susceptibles d'échouer ou d'abandonner, ce qui permet une intervention précise et anticipée [15].

2.4.6. Divertissement

L'intelligence artificielle adapte les suggestions de contenu et améliore l'expérience de l'utilisateur. Des plateformes telles que Netflix et YouTube exploitent des technologies de recommandation alimentées par l'IA pour étudier les habitudes des utilisateurs et suggérer des films, vidéos ou morceaux musicaux en adéquation avec leurs goûts. Dans le domaine du jeu vidéo, l'intelligence artificielle est exploitée pour concevoir des ennemis astucieux qui peuvent évoluer et se modifier en fonction des tactiques des joueurs, proposant une expérience plus vivante et sur mesure [16].

2.4.7. Agriculture

L'intelligence artificielle améliore la gestion des ressources agricoles par le biais du suivi des cultures, de l'estimation des rendements et de la gestion de l'eau. Des drones dotés de capteurs et d'algorithmes d'intelligence artificielle, tels que ceux de PrecisionHawk, procèdent à une surveillance en direct de la santé des cultures, détectant des indices de stress hydrique ou

d'infestation. L'IA contribue aussi à la planification des semis en se basant sur les prévisions météorologiques, ce qui optimise les récoltes [16].

2.4.8. Sécurité

L'intelligence artificielle est de plus en plus mise en œuvre pour assurer une surveillance intelligente et l'identification des menaces. Les systèmes de vidéosurveillance intelligents examinent en direct les séquences vidéo afin d'identifier des comportements irréguliers ou des situations d'urgence. Des instruments tels que Darktrace en cybersécurité emploient l'intelligence artificielle pour repérer d'éventuelles attaques en scrutant des millions d'informations provenant de réseaux informatiques, fournissant ainsi une défense anticipée contre les menaces cybernétiques [16].

2.4.9. Médias et communication

L'intelligence artificielle se distingue particulièrement dans les domaines du traitement du langage naturel (NLP) et de la création de contenu. Les assistants virtuels et chatbots, comme ceux créés par OpenAI ou Google, exploitent l'intelligence artificielle pour communiquer avec les utilisateurs, répondre à leurs interrogations et réaliser des tâches simples. Des systèmes de contrôle automatisé examinent les posts sur les réseaux sociaux afin d'identifier et d'éliminer le contenu inapproprié ou préjudiciable, contribuant ainsi à l'amélioration de la sécurité sur Internet [16].

2.4.10. Commerce de détail

L'intelligence artificielle contribue à optimiser les stocks, à personnaliser l'expérience de l'acheteur et à anticiper les évolutions des ventes. Des systèmes d'IA sont employés par des commerçants tels qu'Amazon pour étudier les habitudes d'achat des clients, ce qui facilite la suggestion de produits sur mesure et l'amélioration de la performance des chaînes d'approvisionnement [16].

2.4.11. Environnement

L'IA est mise en œuvre pour contrôler les écosystèmes, anticiper les désastres naturels et administrer les ressources naturelles. Des systèmes d'intelligence artificielle examinent les données climatiques pour anticiper des phénomènes tels que les inondations ou les feux de forêt, facilitant des interventions plus rapides et performantes. Par ailleurs, l'intelligence artificielle

joue un rôle dans la gestion de la biodiversité en suivant les populations d'espèces à risque et en suggérant des plans de protection [16].

3. Machine Learning

Machine Learning ou encore l'apprentissage automatique, qui est une branche de l'intelligence artificielle, a pour objectif de faire en sorte que les machines puissent apprendre à partir des données, sans nécessiter une programmation explicite. Ces systèmes, grâce à des algorithmes qui détectent des motifs et des tendances, ont la capacité d'améliorer leurs performances progressivement. Au contraire des techniques statistiques traditionnelles, l'apprentissage automatique a la capacité de gérer d'importants volumes de données qu'elles soient structurées ou non (textes, images, etc.) et de représenter des relations complexes généralement non linéaires **[17]**.

Le deep learning ou apprentissage profond est une approche avancée de l'apprentissage automatique qui utilise des réseaux neuronaux artificiels profonds pour apprendre et effectuer des tâches complexes en exploitant de grandes quantités de données. Il a révolutionné de nombreux domaines de l'IA et continue d'être une méthode de pointe pour résoudre des problèmes difficiles liés à la perception, à la compréhension et à la prise de décision [18]. La Figure (1) compare l'apprentissage automatique (Machine Learning - ML) et l'apprentissage profond (Deep Learning - DL).

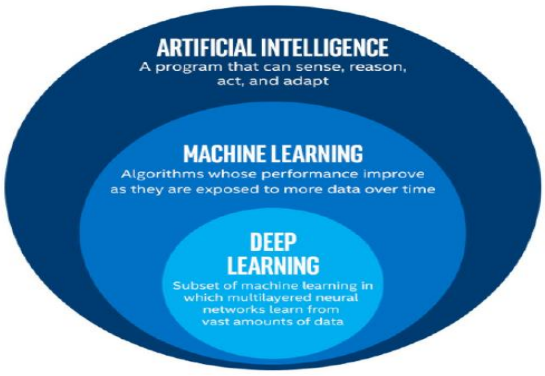


Figure 1 : L'intelligence artificielle, machine Learning et Deep Learning [19]

4. Types d'apprentissage automatique

Dans le domaine de l'apprentissage automatique (ML), on distingue plusieurs formes d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage par renforcement, l'apprentissage par transfert et la figure ci-dessous résume ces types.

4.1. Apprentissage supervisé

Dans cette technique, on forme un modèle sur un ensemble de données labellisées, où les entrées et les sorties sont déjà définies. Le but est de prévoir les résultats pour de nouvelles entrées en se fondant sur les modèles acquis grâce aux données d'apprentissage. Par exemple, des algorithmes tels que les machines à vecteurs de support (SVM) ou les réseaux neuronaux sont fréquemment employés dans le contexte de l'apprentissage supervisé.

L'objectif de l'apprentissage supervisé est d'inférer la relation entre les variables explicatives et la variable à prédire, afin de prédire avec précision la variable à prédire pour de nouvelles entrées pour lesquelles la variable à prédire est inconnue. L'algorithme d'apprentissage supervisé est généralement choisi en fonction de la nature des données et du type de variable à prédire. Les deux types d'algorithmes les plus couramment utilisés en apprentissage supervisé sont les algorithmes de régression et les algorithmes de classification.

- Les algorithmes de régression sont utilisés pour prédire une variable continue, telle que le prix d'une maison ou le nombre de ventes d'un produit.
- Les algorithmes de classification quant à eux, sont utilisés pour prédire une variable catégorielle, telle que le type de fleur ou la catégorie de spam d'un courriel [18].

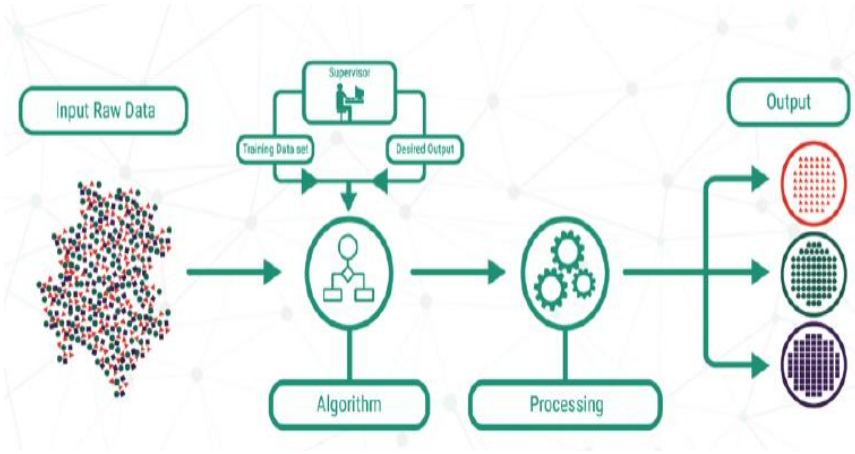


Figure 2 : L'apprentissage supervisé [19]

4.2. Apprentissage non supervisé

À l'inverse de l'apprentissage supervisé, les données utilisées pour l'entraînement ne possèdent pas d'étiquettes. L'intention est d'identifier des structures latentes dans les données, comme la constitution de groupes (clustering) ou l'établissement d'associations. Parmi les

techniques fréquemment employées, on trouve l'algorithme des K-moyennes (K-Means) ou l'analyse en composantes principales (PCA) **[7]**.

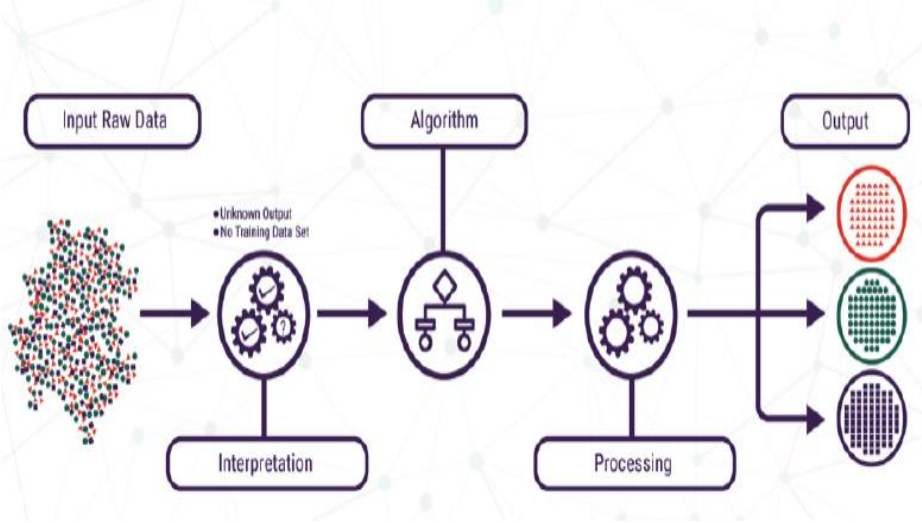


Figure 3 : L'apprentissage non supervisé [19]

4.3. Apprentissage par renforcement

Dans cette approche, un agent se forme à la prise de décisions en interagissant avec un environnement. L'agent obtient des primes ou des sanctions en fonction de ses actes, et il cherche à optimiser la rétribution globale. On recourt souvent à cette technique dans des domaines tels que la robotique et les jeux vidéo, où un agent se doit d'apprendre comment perfectionner son comportement [7].

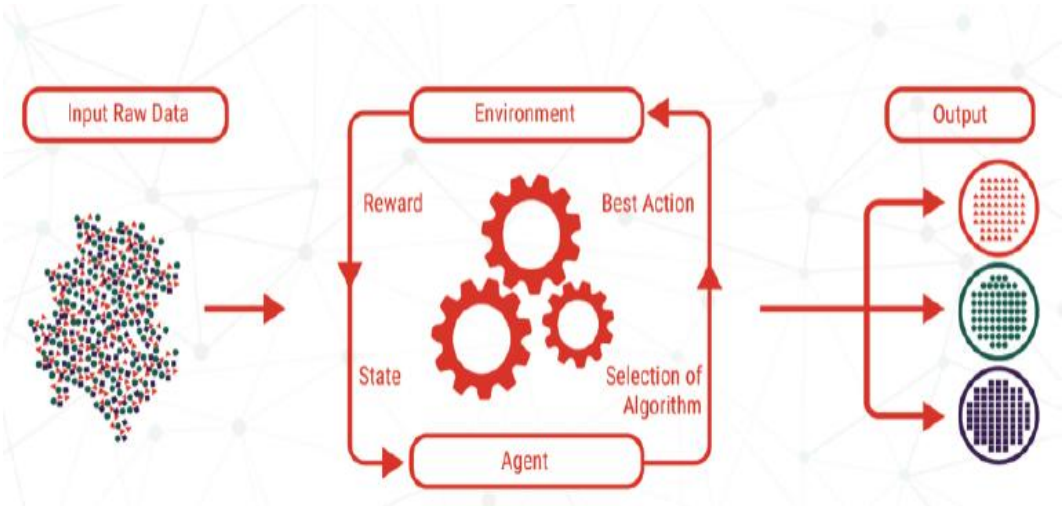


Figure 4 : Apprentissage par renforcement [19]

4.4. Apprentissage semi-supervisé

Cette technique intègre des aspects de l'apprentissage supervisé et non supervisé. Elle s'appuie sur un ensemble réduit de données étiquetées et une vaste collection de données non étiquetées pour optimiser le rendement du modèle. Ce genre d'apprentissage est fréquemment employé lorsque l'annotation des données s'avère onéreuse ou complexe, comme c'est le cas dans la reconnaissance d'images ou le traitement du langage naturel [7].

4.5.Apprentissage par transfert

Cette méthode implique l'utilisation de ce qui a été appris dans une tâche pour optimiser une autre. Elle est particulièrement bénéfique lorsque les informations disponibles pour la tâche nouvelle sont restreintes. Par exemple, des modèles déjà formés sur d'importantes bases de données peuvent être ajustés à de nouvelles tâches en disposant d'un nombre restreint d'exemples, une pratique couramment employée en vision par ordinateur avec les réseaux de neurones convolutifs (CNN).

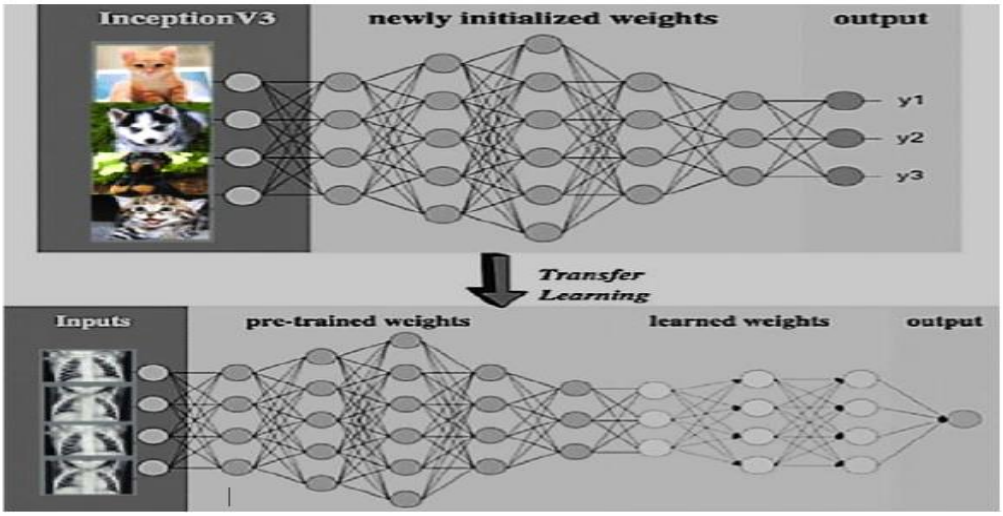


Figure 5 : Apprentissage par transfert [20]

5. Les algorithmes d'apprentissage

5.1. Algorithmes d'apprentissage supervisé

Différents classificateurs d'apprentissage supervisé sont utilisés pour la classification de texte. Voici quelques modèles de base couramment utilisés :

5.1.1. Naïve Bayse

Naïve Bayes est une méthode de classification qui utilise le théorème de Bayes pour calculer des probabilités conditionnelles, en utilisant la règle de Bayes.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad \textbf{(1)}$$

Où :

$P(A|B)$ est la probabilité a posteriori de l'hypothèse A donnée la preuve B. C'est la probabilité que nous cherchons à déterminer.

$P(B|A)$ est la probabilité de la preuve B étant donné que l'hypothèse A est vraie. $P(A)$ est la probabilité a priori de l'hypothèse A, qui est notre croyance sur la probabilité de A avant de considérer la preuve B.

$P(B)$ est la probabilité totale de l'évidence B, souvent appelée la vraisemblance marginale ou la constante de normalisation. Cela représente la probabilité d'observer la preuve B indépendamment de la véracité de l'hypothèse A.

La règle de Bayes, également connue sous le nom de théorème de Bayes, peut être exprimée à l'aide de la formule suivante :

$$P(y|x1, x2, .., xn) = \frac{P(y) \times P(x1|y) \times P(x2|y) \times ... \times P(xn|y)}{P(x1) \times P(x1) \times .. P(xn)} \quad \textbf{(2)}$$

Où :

$P(y|x1, x2, .., xn)$ est la probabilité a posteriori de la classe y donnée les caractéristiques $x1, x2, .., xn$.

$P(y)$ est la probabilité a priori de la classe y.

$P(xi|y)$est la vraisemblance de la caractéristique xi donnée la classe y. $P(xi)$ est la probabilité de la caractéristique $x1, x2, .., xn$ sont les caractéristiques.

5.1.2. Machines à vecteurs de support

Les machines à vecteurs de support (SVM) sont des algorithmes de classification conçus pour trouver un classificateur optimal qui sépare efficacement les points de données et maximise la marge entre deux classes. Ce classificateur est représenté par un hyperplan linéaire. Le plan hyper divise les points de données en deux ensembles. Les points de données les plus proches de l'hyperplan, et cruciaux pour déterminer sa position, sont appelés vecteurs de support.

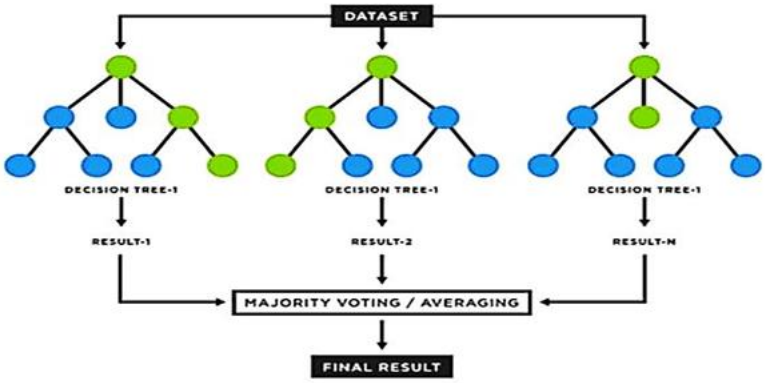


Figure 6 : Machines à vecteurs de support [20]

5.1.3. Random Forests

Random Forests sont une implémentation d'algorithmes basés sur des arbres de décision qui permettent de modéliser les résultats en fonction des choix précédents le long de différentes branches. En considérant plusieurs arbres de décision, il vise à prendre la décision optimale en fonction des résultats ultérieurs. Cette approche peut être considérée comme une forme d'anticipation, où les prédictions collectives de plusieurs arbres contribuent à un résultat plus complet et plus fiable.

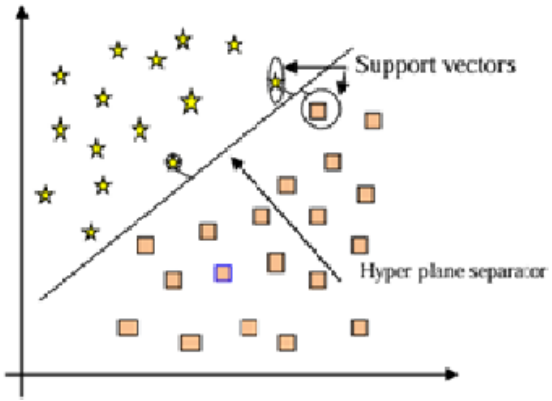


Figure 7 : Représentation graphique des forêts aléatoires [7]

5.1.4. KNN (k-Nearest Neighbors)

KNN (k-Nearest Neighbors) est un algorithme de classification qui attribue de nouveaux points de données à des classes en fonction de la majorité de leurs points de données voisins. Dans

l'algorithme des k-Plus Proches Voisins (KNN), les distances entre les points de données sont cruciales pour déterminer la "proximité" ou la similarité entre les instances.

5.1.5. Arbre de décision

Un est un outil précieux utilisé pour les problèmes de classification, utilisant une structure ressemblant à un organigramme. Chaque nœud interne de l'arbre de décision représente une condition ou un "test". Basé sur un attribut, l'arbre se ramifie en conséquence. Les nœuds feuilles de l'arbre contiennent des étiquettes de classe, qui sont déterminées après l'évaluation de tous les attributs. Le chemin de la racine à un nœud feuille représente une règle de classification.

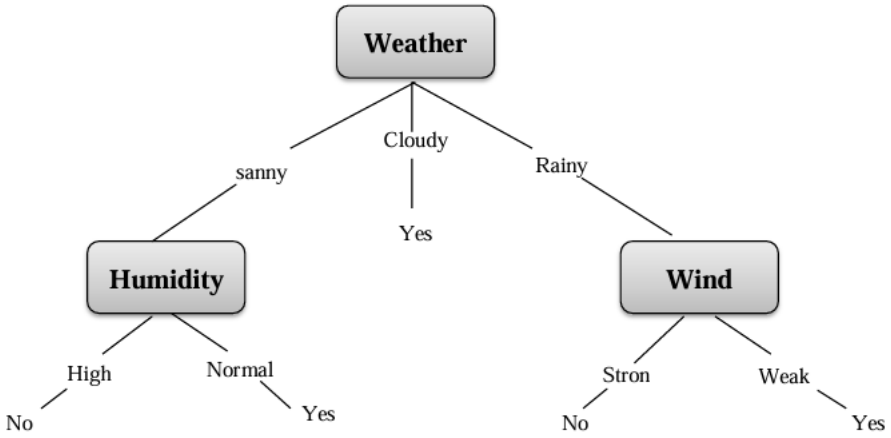


Figure 8 : Représentation graphique de l'Arbre de Décision [7]

5.2. Algorithmes d'apprentissage non supervisé

Les algorithmes d'apprentissage non supervisé sont mieux adaptés aux problèmes plus complexes. Ci-dessous, nous décrivons quelques exemples.

6. L'algorithme K-means

L'algorithme des k-moyennes (ou K-means en anglais) est un algorithme de partitionnement de données relevant des statistiques et de l'apprentissage automatique (plus précisément de l'apprentissage non supervisé), le plus connu et le plus utilisé, du fait de sa simplicité de mise en œuvre. C'est une méthode dont le but est de diviser des observations en K partitions (clusters) dans lesquelles chaque observation appartient à la partition avec la moyenne la plus proche [21].

6.1. Principe de la méthode des K-Means

Le principe du K-means est de segmenter les données en k-groupes. La Figure (9) ci-dessous présente un exemple explicatif de la mise en œuvre de l'algorithme K-Means, comprenant trois regroupements (voir figure 9 ci-aprés) [22].

- Dans un premier temps, k points sont sélectionnés de manière semi-aléatoire pour servir de centres aux clusters. Chaque instance est attribuée au centre qui lui est le plus proche, cette proximité étant déterminée par la distance euclidienne.
- Par la suite, les centres de chaque regroupement formé sont réévalués en fonction de l'emplacement des instances qu'ils englobent. Ensuite, les instances sont attribuées à chaque cluster selon leur éloignement euclidien par rapport aux centres récemment déterminés.
- Enfin, cette procédure est répétée jusqu'à ce que les centres des clusters subissent à peine des variations d'une itération à l'autre, un phénomène que l'on désigne comme la stabilisation des centres de gravité. Par conséquent, on génère des regroupements qui ne se chevauchent pas et qui englobent l'ensemble des instances de la base de données.

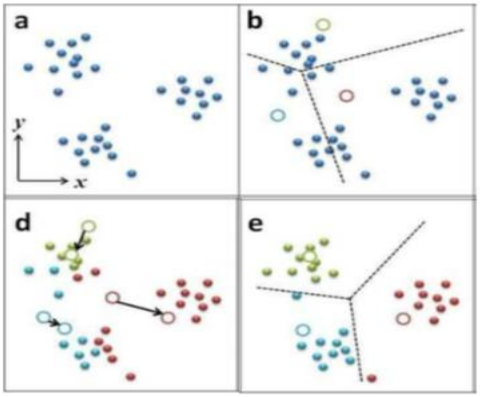


Figure 9 : Principe de l'algorithme K-means **[22]**

6.2. Fonctionnement de l'algorithme K-means

L'algorithme suivant résume les étapes principales pour l'exécution [23] :

Entrée
Ensemble de N données, noté par $X = \{x1, x2, .. xn\}$
Nombre de groupes souhaité, noté par k
Sortie
Une partition de K groupes ***{C1,C2, ... C_k}***
Début
1) Initialisation aléatoire des centres C_k

Répéter
2) Etape d'affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche

$$x_i \in C_k \quad \text{si} \quad \|x_i - \mu_k\| = \min_{1 \le j \le k} \|x_i - \mu_j\|$$

3) Etape de mise à jour : Calculer les centres associe à la nouvelle partition

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

Jusqu'à convergence de l'algorithme vers une partition stable

6.3. Convergence et initialisation des K-Means

On définit la convergence comme un minimum local de l'énergie, qui se manifeste par la division de l'espace des données en classes distinctes à travers des hyperplans. La qualité de la solution obtenue est fortement influencée par les noyaux de départ. Par ailleurs, l'algorithme est d'autant plus sensible à l'initialisation lorsque la dimensionnalité des données est élevée [24].

6.4. Choix du nombre K de classes (clusters)

L'algorithme classique des K-means permet une certaine flexibilité : le choix du nombre de classes (ou clusters) à constituer. Dans le contexte de l'examen des données sous forme de tableau, ce paramètre illustre le nombre de groupes homogènes que nous visons à reconnaître dans l'ensemble de données. Pour obtenir une segmentation significative, le choix approprié de K est crucial. Des approches telles que la méthode du coude ou le **score de silhouette** peuvent contribuer à trouver une valeur optimale pour K [24].

9. Conclusion

Ces principes fondamentaux trouvent leur application directe dans le chapitre suivant, qui présente une étude de cas réelle de segmentation client via le clustering K-Means. La théorie devient ainsi un levier d'action pour le marketing personnalisé.

Chapitre 3
Segmentation Client par Clustering K-Means

1. Introduction

Dans un environnement commercial concurrentiel, la compréhension des besoins clients constitue un avantage décisif. Cette étude de cas applique l'algorithme K-Means à la segmentation d'une clientèle de centre commercial disposant d'un programme de fidélité. L'objectif consiste à transformer des données clients brutes en stratégies marketing personnalisées grâce aux techniques d'apprentissage automatique non supervisé [25]. La démarche progresse logiquement de l'analyse exploratoire vers des recommandations stratégiques opérationnelles. Cette approche scientifique illustre la valeur ajoutée de l'intelligence artificielle appliquée au marketing relationnel.

2. Étude de Cas : Segmentation Client par Clustering K-Means

Dans un contexte où la personnalisation client devient stratégique, l'analyse comportementale permet aux centres commerciaux d'identifier des segments de consommateurs aux habitudes similaires. Cette approche data-driven utilise des algorithmes de clustering pour transformer les données clients en insights stratégiques, optimisant ainsi l'offre commerciale et les stratégies marketing selon les profils identifiés [26].

2.1. Contexte et Problématique

2.1.1. Présentation du cas d'étude

Un centre commercial souhaite optimiser sa stratégie marketing en segmentant sa clientèle. Grâce aux données collectées via son programme de fidélité, l'établissement dispose d'informations comportementales sur 200 clients : âge, genre, revenu annuel et score de dépenses.

2.1.2. Objectifs de l'étude

- **Objectif principal** : Identifier des segments de clients homogènes pour personnaliser l'offre commerciale
- **Objectifs secondaires**
    - Comprendre les profils de consommation selon l'âge et le revenu
    - Optimiser l'allocation budgétaire marketing par segment
    - Développer des stratégies ciblées par typologie de clients

3. Méthodologie retenue

Nous allons réaliser la segmentation des clients du centre commercial par l'algorithme de clustering non supervisé **K-Means**. Cette segmentation sera précéder par des statistiques descriptives à fin de mieux comprendre le jeu de données que nous allons employé.

3. 1. Analyse Exploratoire des Données

Le jeu de données employé dans cette étude a été téléchargé à partir du lien [27] du site Kaggle, il est constitué d'un seul fichier **Mall_Customers.csv**.

3.1.1. Résumé de la Qualité des Données (Data Quality Summary)

Avant de procéder aux différents traitements, nous allons présenter le jeu de données que nous avons sélectionné dans cette étude. La présentation du jeu de données permet de le comprendre mieux. Le tableau (1) ci-après montre les enregistrements qui constituent le jeu de donnée Mall_Customers employé.

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Tableau 1 : Le jeu de données Mall_Customers

Le jeu de données contient 5 colonnes et 200 lignes. Nous résumons ses caractéristiques en :

- Taille du dataset : 200 observations × 5 variables
- Complétude : 100% (aucune valeur manquante)
- Unicité : 100% (aucun doublon détecté)
- Cohérence : Variables bien définies et cohérentes

3.1.2. Profil des variables

Le tableau (2) ci-après, décrit les profils des variables

Variable	Type	Valeurs Uniques	Min	Max	Valeurs Manquantes
CustomerID	Numérique	200	1	200	0
Gender	Catégorielle	2	-	-	0
Age	Numérique	47	18	70	0
Annual Income (k$)	Numérique	68	15	137	0
Spending Score (1-100)	Numérique	73	1	99	0

Tableau 2 : Profil des variables du jeu de données Mall_Customers

3.2. Statistiques Descriptives

Avant d'effectuer le clustering basé K-means, nous avons effectué des calculs statistiques sur le jeu de données, pour comprendre leur distribution, les mesures de tendance centrale et leurs dispersion les calculs couvent : la moyenne, le médiane, l'écart-type, la variance…etc. (Voir tableau (3)).

Métrique	Âge	Revenu Annuel (k$)	Score de Dépenses
Moyenne	38,85	60,56	50,20
Médiane	36,00	61,50	50,00
Mode	32	54	42
Écart-type	13,97	26,26	25,82
Variance	195,13	689,59	666,69
Minimum	18	15	1
Maximum	70	137	99

Tableau 3 : Statistiques descriptives sue le jeu de données Mall_Customers

Interprétation des statistiques :

- **Âge** : Distribution légèrement asymétrique vers la droite (moyenne > médiane)
- **Revenu** : Distribution relativement symétrique autour de 60k$
- **Score de dépenses** : Distribution uniforme sur l'échelle 1-100

3.3. Analyse par Genre

Le tableau (4) ci-après, présente la répartition de la clientèle par genre

Genre	Effectif	Pourcentage	Proportion
Féminin	112	56,0%	0,56
Masculin	88	44,0%	0,44
Total	200	100,0%	1,00

Tableau 4 : Répartition de la clientèle par genre

3.4. Statistiques descriptives par genre

Le tableau (5) ci-après, présente les statistiques descriptives par genre

Métrique	Femmes	Hommes	Différence
Âge moyen	38,1 ans	39,8 ans	-1,7 ans
Revenu moyen	59,3 k$	62,2 k$	-2,9 k$
Score dépenses moyen	51,5	48,5	+3,0

Tableau 5 : Statistiques descriptives par genre

Observations clés :

- Léger déséquilibre en faveur de la clientèle féminine (56%)
- Les hommes ont un revenu légèrement supérieur (+4,9%)
- Les femmes présentent un score de dépenses plus élevé (+6,2%)

3.5. Analyse par Groupe d'Âge

Le tableau (5) ci-après, présente la segmentation par tranches d'âge

Groupe d'Âge	Plage	Effectif	Pourcentage	Âge Moyen	Revenu Moyen	Score Moyen
Jeunes	18-30 ans	56	28,0%	25,4	56,2	49,1
Adultes	31-50 ans	92	46,0%	41,1	61,5	50,8
Seniors	51-70 ans	52	26,0%	59,2	63,2	50,8

Tableau 6 : La segmentation par tranches d'âge

Profil comportemental par groupe d'âge :

- **Jeunes (18-30 ans)** :
  - Revenus légèrement inférieurs à la moyenne (-7,2%)
  - Comportement de dépense proche de la moyenne
  - Potentiel de croissance du pouvoir d'achat
- **Adultes (31-50 ans)** :
  - Groupe le plus représenté (46% de la clientèle)
  - Revenus supérieurs à la moyenne (+1,6%)
  - Stabilité dans les habitudes de consommation
- **Seniors (51-70 ans)** :
  - Revenus les plus élevés (+4,4% vs moyenne)
  - Comportement de dépense stable
  - Fidélité présumée plus importante

3.6. Analyse Croisée Âge-Genre

Le tableau (7) ci-après, présente la répartition croisée âge-genre

Groupe d'Âge	Femmes	% Femmes	Hommes	% Hommes	Total
Jeunes (18-30)	32	57,1%	24	42,9%	56
Adultes (31-50)	53	57,6%	39	42,4%	92
Seniors (51-70)	27	51,9%	25	48,1%	52

Tableau 7 : La répartition croisée âge-genre

3.7. Analyse comportementale détaillée

Le tableau (8) ci-après, présente une analyse comportementale du jeu de données Mall_Customers.

Segment	Effectif	Âge Moyen	Revenu Moyen	Score Moyen
Femmes Jeunes	32	25,1	53,8	50,2
Hommes Jeunes	24	25,8	59,7	47,4
Femmes Adultes	53	40,8	58,9	52,1
Hommes Adultes	39	41,5	65,1	48,9
Femmes Seniors	27	59,1	62,0	52,3
Hommes Seniors	25	59,3	64,5	49,2

Tableau 8 : Analyse comportementales

Recommandations stratégiques :

- **Constante de genre** : Les femmes maintiennent un score de dépenses supérieur dans tous les groupes d'âge
- **Évolution du revenu** : Progression naturelle avec l'âge, plus marquée chez les hommes
- **Potentiel par segment** :
    - Femmes adultes : segment le plus important (26,5% de la clientèle)
    - Hommes jeunes : potentiel de développement du score de dépenses
    - Seniors : stabilité et pouvoir d'achat élevé

4. La segmentation des clients par K-means

4. 1. Détermination du Nombre Optimal de Clusters

Pour déterminer le meilleur nombre de cluster pour l'algorithme K-means, nous effectuant une analyse comparative des Solutions de Clustering en variant la valeur de k [29]. Pour notre cas nous avons choisi les valeurs k=3, k =4, k=5 et k=6.

4.1.1. Segmentation Basique avec k = 3

3. **Profils identifiés :**

- **Cluster 0** : *Jeunes Économes* - Faible revenu, dépenses modérées
- **Cluster 1** : *Jeunes Aisés Dépensiers* - Revenu élevé, forte propension à la consommation
- **Cluster 2** : *Seniors Prudents* - Âge avancé, dépenses limitées

4. **Avantages :** Simplicité d'interprétation
5. **Inconvénients :** Perte de nuances comportementales importantes

4.1.2. Segmentation Intermédiaire avec k = 4

Profils enrichis :

- **Cluster 0** : *Quadragénaires Aisés Conservateurs* - Revenu élevé mais dépenses très faibles
- **Cluster 1** : *Trentenaires Affluents* - Revenu et dépenses très élevés
- **Cluster 2** : *Jeunes Consommateurs Actifs* - Faible revenu mais dépenses élevées
- **Cluster 3** : *Seniors à Budget Limité* - Revenus et dépenses modérés

4.1.5. Segmentation Intermédiaire avec k = 5

Profils encore enrichis

- **Cluster 0 :** Personnes âgées avec un revenu moyen et un score de dépenses moyen.
- **Cluster 1 :** Personnes de 30 ans avec un revenu élevé et un score de dépenses très élevé.
- **Cluster 2 :** Jeunes avec un faible revenu et un score de dépenses moyen-élevé.
- **Cluster 3 :** Personnes d'environ 50-60 ans avec un très faible revenu et un score de dépenses très faible.
- **Cluster 4 :** Personnes d'environ 40 ans avec un revenu élevé et un score de dépenses très faible.

4.1.3. Segmentation optimale avec k = 6

Le tableau (9) ci-après, présente le résultat de la segmentation k-means du jeu de données Mall_Customers, avec k=6.

Cluster	Profil	Âge Moyen	Revenu	Score Dépenses	Stratégie Recommandée
0	*Jeunes Impulsifs*	25-30	Très faible	Très élevé	Produits tendance
1	*Seniors Équilibrés*	50-60	Moyen	Moyen	Produits qualité-prix, services
2	*Cadres Épargnants*	40-45	Très élevé	Très faible	Produits premium, investissements
3	*Jeunes Modérés*	25-35	Moyen	Moyen	Promotions, gamme moyenne
4	*Profils Affluents*	30-35	Élevé	Très élevé	Produits haut de gamme, expériences
5	*Seniors Précaires*	50-55	Très faible	Très faible	Offres économiques, promotions

Tableau 9 : Analyse comportementales

4.2 Score de Silhouette [28] [29]

Le score de silhouette confirme que **k = 6** produit des clusters bien séparés et cohérents, avec des valeurs également acceptables pour k = 9 et k = 10 (figure 10)

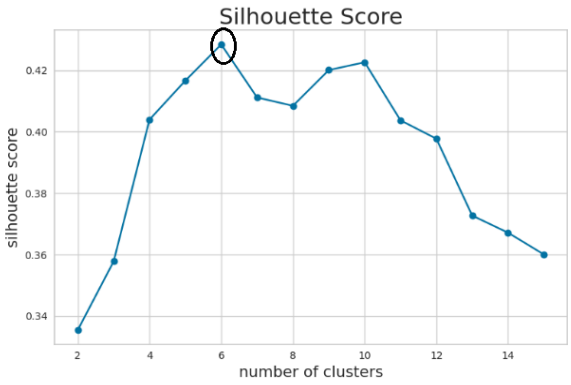


Figure 10 : Le score de silhouette

4.3. Construction de personas

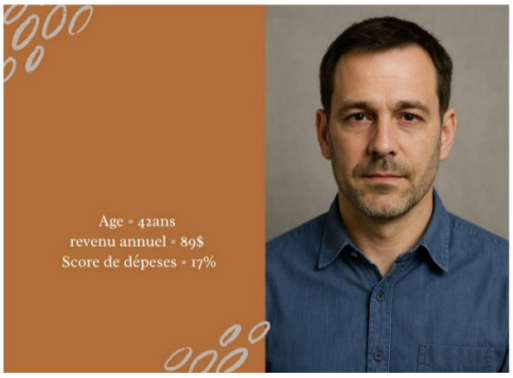
L'application de l'algorithme K-Means avec k = 6 clusters révèle une segmentation client pertinente pour le centre commercial. Cette typologie permet de différencier clairement les comportements d'achat selon l'âge et le niveau de revenu, offrant une base solide pour des stratégies marketing personnalisées. Nous allons construis les personas relatives aux segments clients.

Leila

Revenu très faible et dépense très élevé

Asma

Revenu très faible et dépense très élevé



Mohamed

Revenu très élevé et dépenses très faibles

Donia

Revenu moyen et dépenses moyen

Hakim

Revenu élevé et dépenses très élevé

Ikram

Revenu très faible et dépenses très faible

5. Recommandations Stratégiques

5.1 Stratégies Marketing par Segment

- **Pour les Jeunes Impulsifs : persona Leila**
- Campagnes sur réseaux sociaux
- Promotions flash et offres limitées
- Partenariats avec influenceurs

✓ **Pour les Cadres Épargnants : persona Mohamed**

- Communication axée sur la valeur et la durabilité
- Programmes de fidélité à long terme
- Produits d'investissement et épargne

✓ **Pour les Profils Affluents : persona Hakim**

- Services personnalisés
- Communication premium multicanal

6. Validation et Limites de l'Étude

6.1 Forces de l'analyse

- Dataset complet sans valeurs manquantes
- Méthodes de validation convergentes (indice de silhouette)
- Interprétation business claire des segments

6.2 Limites identifiées

- Taille d'échantillon limitée (200 observations)
- Variables comportementales restreintes
- Absence de données temporelles (évolution des comportements)

6.3 Recommandations pour l'amélioration

- Enrichissement du dataset avec données transactionnelles détaillées
- Intégration de variables saisonnières
- Mise en place d'un système de re-segmentation périodique

7. Conclusion

Cette étude a démontré l'efficacité du clustering K-Means pour segmenter une clientèle de centre commercial en six profils distincts, validés par des méthodes statistiques robustes. L'analyse a révélé des segments allant des "Jeunes Impulsifs" aux "Cadres Épargnants", chacun nécessitant des stratégies marketing spécifiques.

Cette approche data-driven constitue un outil stratégique incontournable pour développer une relation client personnalisée dans un environnement concurrentiel. Les perspectives d'évolution incluent l'intégration de données temps réel et l'extension vers des techniques de machine learning supervisé pour une optimisation continue des campagnes marketing.

Conclusion générale

Cette étude a démontré l'efficacité de l'algorithme K-means pour segmenter une clientèle de centre commercial en groupes homogènes présentant des comportements d'achat distincts. L'analyse des données démographiques et financières a permis d'identifier six profils clients clairs, allant des "Jeunes Impulsifs" aux "Cadres Épargnants", chacun nécessitant des stratégies marketing spécifiques.

Sur le plan méthodologique, l'approche a combiné analyse exploratoire (statistiques descriptives, visualisations) et techniques de clustering (score de silhouette) pour valider la robustesse des segments. L'étude met en lumière plusieurs enseignements pratiques :

1) L'importance des variables démographiques de base (âge, revenu) pour une première segmentation

2) La valeur ajoutée du croisement entre données financières et comportementales

3) La nécessité d'adapter les canaux de communication à chaque segment (réseaux sociaux pour les jeunes, approche premium pour les clients aisés)

Les limites principales concernent :

- La sensibilité du K-means à l'initialisation aléatoire
- Le choix parfois subjectif du nombre de clusters
- L'absence de données temporelles pour analyser l'évolution des comportements

Pour les professionnels du marketing, cette recherche propose un cadre opérationnel immédiatement applicable :

• Création de personas basés sur les clusters identifiés
• Allocation différenciée du budget marketing
• Personnalisation des promotions et offres produits

Les perspectives de recherche suggèrent notamment :

1) Le test d'algorithmes alternatifs (DBSCAN, clustering hiérarchique)

2) L'intégration de données complémentaires (historique d'achats, canaux digitaux)

3) L'hybridation avec des techniques supervisées pour prédire l'appartenance aux segments

En conclusion, cette approche data-driven transforme la relation client d'une logique de masse vers un marketing ultra-personnalisé. Les centres commerciaux disposent désormais d'une méthodologie éprouvée pour optimiser leurs investissements marketing tout en améliorant l'expérience client. Les prochains chapitres exploreront comment enrichir cette segmentation grâce à l'analyse prédictive et l'IA générative.

Bibliographie

[1] A. &. S. J. N. Parvatiyar, «Customer Relationship Management: Emerging Practice, Process, and Discipline. Journal of Economic & Social Research, 3(2).,» 2001.

[2] W. R. Smith, «Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing, 21(1), 3-8.,» 1956.

[3] N. Chleq, «Marketing stratégique. Pearson Éducation.,» 2011.

[4] P. &. K. K. L. Kotler, «Marketing Management (13e éd.). Pearson Education.,» 2009.

[5] S. &. N. P. Russell, «Artificial Intelligence: A Modern Approach. Pearson.,» (2021).

[6] I. B. Y. &. C. A. Goodfellow, «Deep Learning. MIT Press.,» 2016.

[7] C. M. Bishop, « Pattern Recognition and Machine Learning. Springer.,» 2006.

[8] H. Tan, «A brief history and technical review of the expert system research. IOP Conference Series.,» *2017.*

[9] R. M. R. M. S. A. &. S. P. J. Priyadarshini, «Artificial Intelligence: Applications and Innovations. CRC Press.,» 2022.

[10] D. &. A. L. O. Baidoo-Anu, «Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Available at SSRN 4337484.,» 2023.

[11] Q. e. a. Bi, « What is machine learning? A primer for the epidemiologist. American Journal of Epidemiology, 188(12), 2222-2239.,» 2019.

[12] M. Krichen, « Amélioration des performances de l'Intelligence Artificielle.,» 2024.

[13] F. W. Y. &. H. G. Yu, «Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. IEEE Geoscience and Remote Sensing Magazine, 6(2). doi: 10.1109/MGRS.2016.2540798.,» 2018.

[14] B. &. K. O. Siciliano, Springer Handbook of Robotics. Springer., 2016.

[15] J. e. a. Schleiss, AI course design planning framework: Developing domain-specific AI education courses. Education Sciences, 13(9), 954., 2023.

[16] P. Zaraté, L'intelligence artificielle d'hier à aujourd'hui. Droit Social, Dalloz, C2., 2021.

[17] R. e. a. Sun, Intelligence artificielle en radiothérapie : radiomique, pathomique, et prédiction de la survie et de la réponse aux traitements. Cancer/Radiothérapie, 25(6-7), 630-637., 2021.

[18] Z. &. O. A. Mimoune, Développement d’une Architecture, 2019.

[19] A. Balodi, «Application of Introduction Artificial Intelligence Machine Learning in Real Life,» April 2020.

[20] M. N. HATEM Rayane, «Combinaison emojis-texte dans une architecture,» Département d'informatique, Université 20 Aout 1955, 2023.

[21] E. K. Brahim, «Etude de méthodes de Clustering pour la segmentation».

[22] B. O. L. Houria, «Classification Des Feuilles de Plantes à Base de Moment de He,» 2014.

[23] L. J. B.SC., «Forage non supervisé de données pour la prédiction d’activité dans les,» Juin2013..

[24] K. Mokran, « i, « Segmentation d'images par classiffeurs non supervisés : Application à l'Imagerie par Résonance Magnétique (IRM) »,Mémoire de Magister, Université A.Mira de Bejaia.2008.,» 2008.

[25] Y. e. a. Liu, «"AI-powered customer segmentation in retail: A K-means clustering case study". Expert Systems with Applications,» (2022).

[26] T. Davenport, «"The AI Advantage in Retail: How Algorithms Are Reshaping Customer Experience". MIT Sloan Management Review.,» 2023.

[27] [En ligne]. Available: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python. [Accès le 15 Mai 2025].

[28] A. Kassambara, «"Practical Guide to Cluster Analysis in R". STHDA.,» *2020.*

[29] M. e. a. Charrad, «"NbClust 3.0: An R Package for Determining the Optimal Number of Clusters". Journal of Statistical Software.,» (2023).

[30] S. e. a. Gupta, « "Behavioral Clustering for Targeted Marketing: A Case Study in Retail". Journal of Retailing.,» 2021.

Annexe

Environnement et outils développement

Introduction

Cette annexe présente les outils matériels et logiciels essentiels utilisés pour implémenter notre système. Elle couvre les composants matériels et logiciels clés qui répondent aux besoins informatiques.

Le matériel

Nom de l'appareil DESKTOP-SFA0DAG
Processeur Intel(R) Core(TM) i7-5600U CPU @ 2.60GHz 2.59 GHz
Mémoire RAM installée 8,00 Go (7,88 Go utilisable)
ID de périphérique 0836C993-031B-4071-9F57-C9A2029DEF5E
ID de produit 00331-10000-00001-AA630
Type du système Système d'exploitation 64 bits, processeur x64
Stylet et fonction tactile La fonctionnalité d'entrée tactile ou avec un stylet n'est pas disponible sur cet écran
Édition Windows 10 Professionnel,version 22H2, installé le 14/09/2021
Build du système d'exploitation 19045.5854
Expérience Windows Feature Experience Pack 1000.19061.1000.0

Bibliothèques Python

pandas
numpy
matplotlib
seaborn
scikit-learn
scipy
yellowbrick

Environment

Visual Studio Code, PyCharm

Comprehensive Data Summary

Dataset Overview

Metric	Value
Total Records	200
Features	5
Numeric Features	4
Categorical Features	1

Data Quality Summary

Feature	Type	Missing	Missing %
CustomerID	int64	0	0.0
Gender	object	0	0.0
Age	int64	0	0.0
Annual Income (k$)	int64	0	0.0
Spending Score (1-100)	int64	0	0.0

Descriptive Statistics

	Age	Annual Income (k$)	Spending Score (1-100)
count	200.0	200.0	200.0
mean	38.85	60.56	50.2
std	13.97	26.26	25.82
min	18.0	15.0	1.0
25%	28.75	41.5	34.75
50%	36.0	61.5	50.0
75%	49.0	78.0	73.0
max	70.0	137.0	99.0

Gender Analysis

	Mean Age	Count	Mean Income	Mean Spending
Female	38.1	112.0	59.25	51.53
Male	39.81	88.0	62.23	48.51

Age Group Analysis

	Mean Income	Count	Mean Spending
18-31	55.21	62.0	61.03
31-44	69.21	67.0	54.4
44-57	59.86	43.0	35.65
57-70	52.79	28.0	38.5

Statistical Quality Metrics

Feature	Outliers %	Skewness
Age	0.0%	0.48
Annual Income (k$)	1.0%	0.32
Spending Score (1-100)	0.0%	-0.05

Categorical Features Distribution

Feature	Value	Count	Percentage
Gender	Female	112	56.0%
Gender	Male	88	44.0%
ageGroups	31-44	67	33.5%
ageGroups	18-31	62	31.0%
ageGroups	44-57	43	21.5%
ageGroups	57-70	28	14.0%

Dataset Overview

Metric	Value
Total Records	200
Features	5
Numeric Features	4
Categorical Features	1

Data Quality Summary

Feature	Type	Missing	Missing %
CustomerID	int64	0	0.0
Gender	object	0	0.0
Age	int64	0	0.0
Annual Income (k$)	int64	0	0.0
Spending Score (1-100)	int64	0	0.0

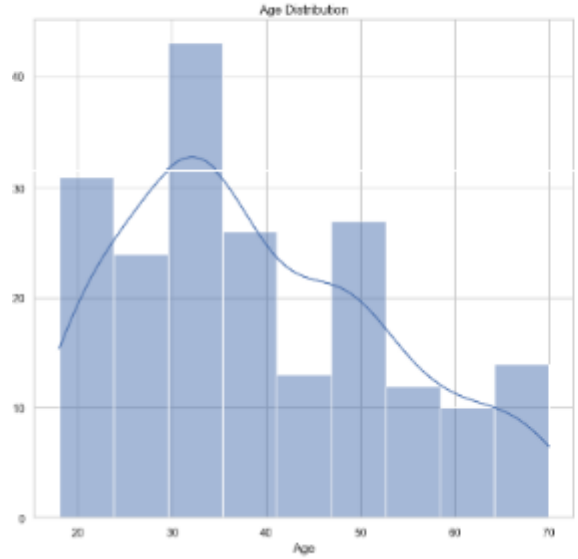
Descriptive Statistics

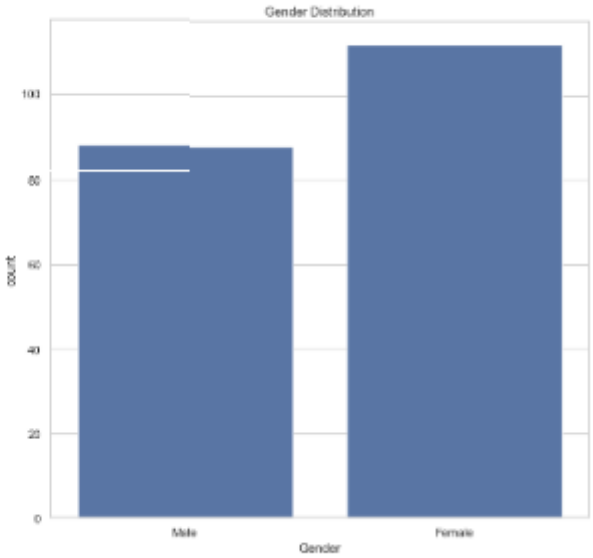
	Age	Annual Income (k$)	Spending Score (1-100)
count	200.0	200.0	200.0
mean	38.85	60.56	50.2
std	13.97	26.26	25.82
min	18.0	15.0	1.0
25%	28.75	41.5	34.75
50%	36.0	61.5	50.0
75%	49.0	78.0	73.0
max	70.0	137.0	99.0

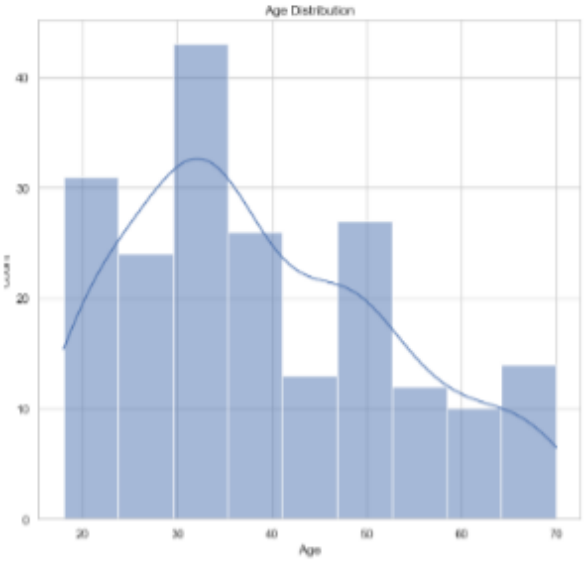
Gender Analysis

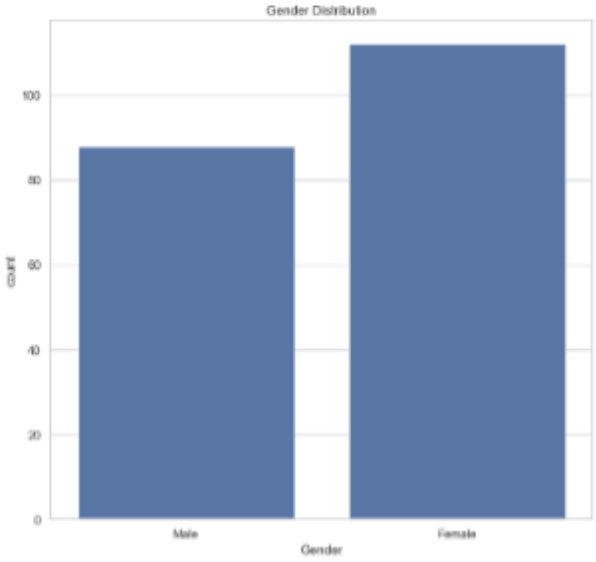
	Mean Age	Count	Mean Income	Mean Spending
Female	38.1	112.0	59.25	51.53
Male	39.81	88.0	62.23	48.51

Age and Gender Distributions

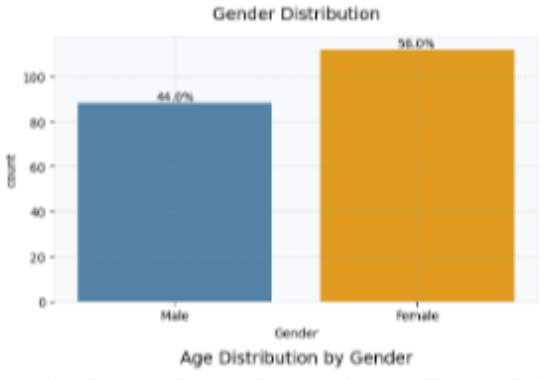




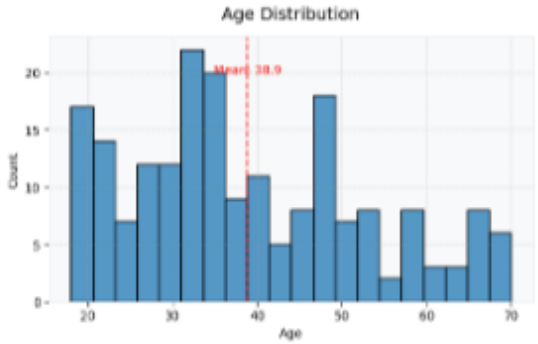




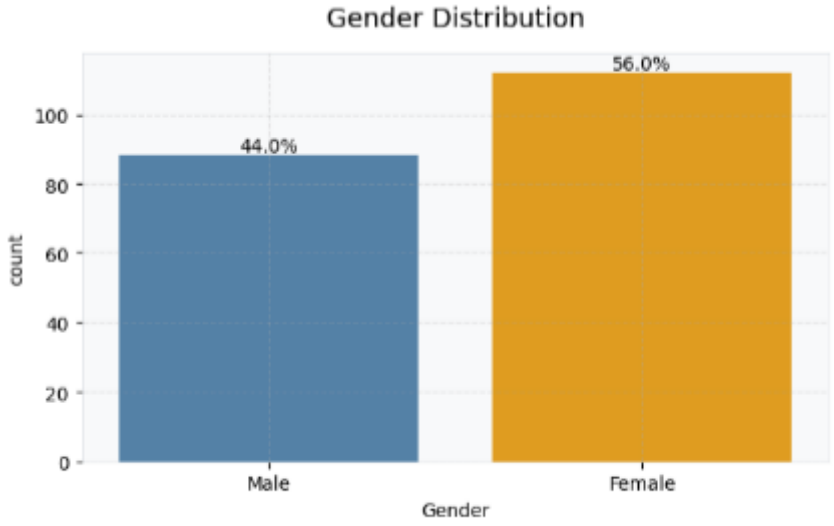
Age and Gender Analysis

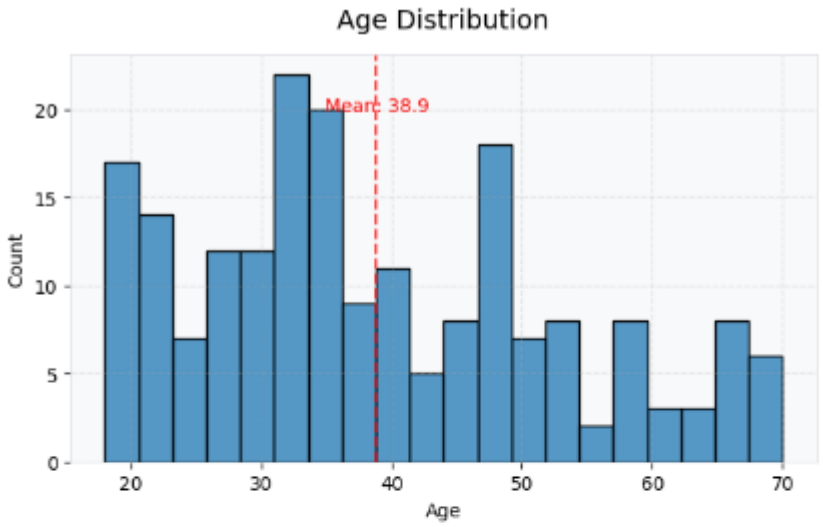


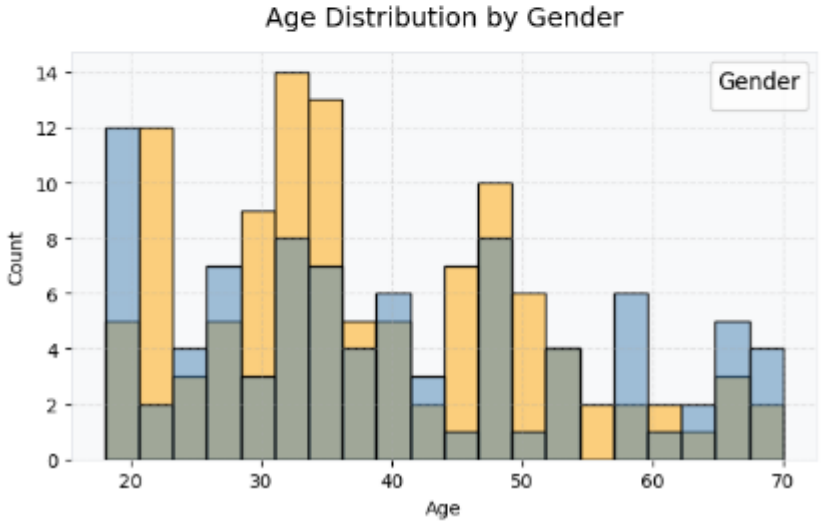
Age Distribution by Gender



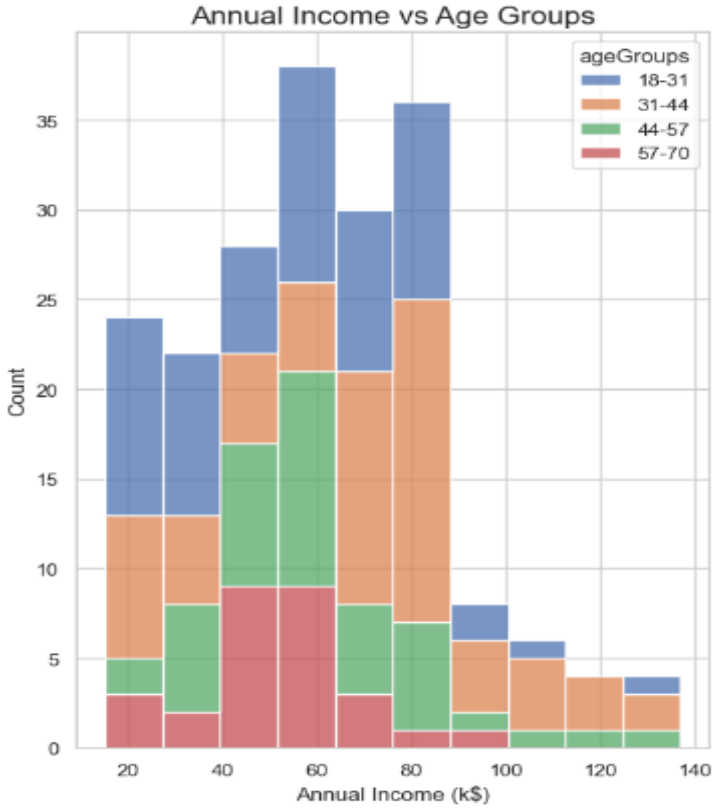
Age Statistics by Gender

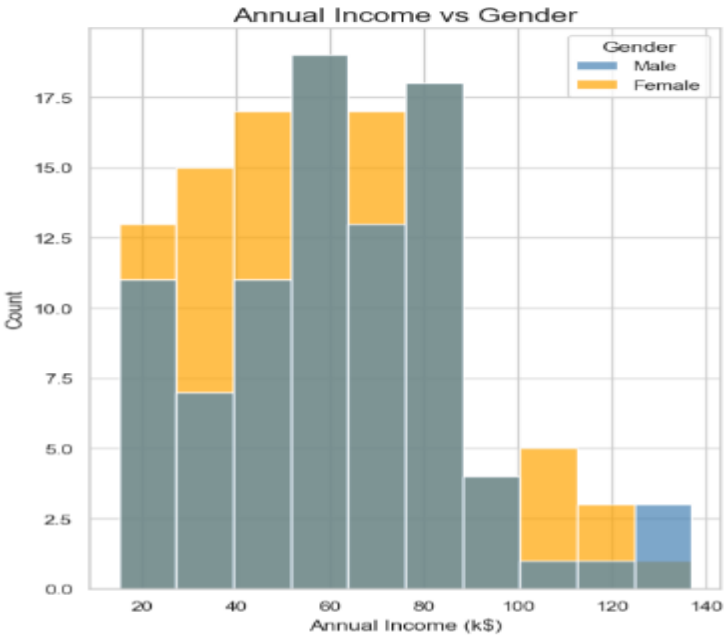


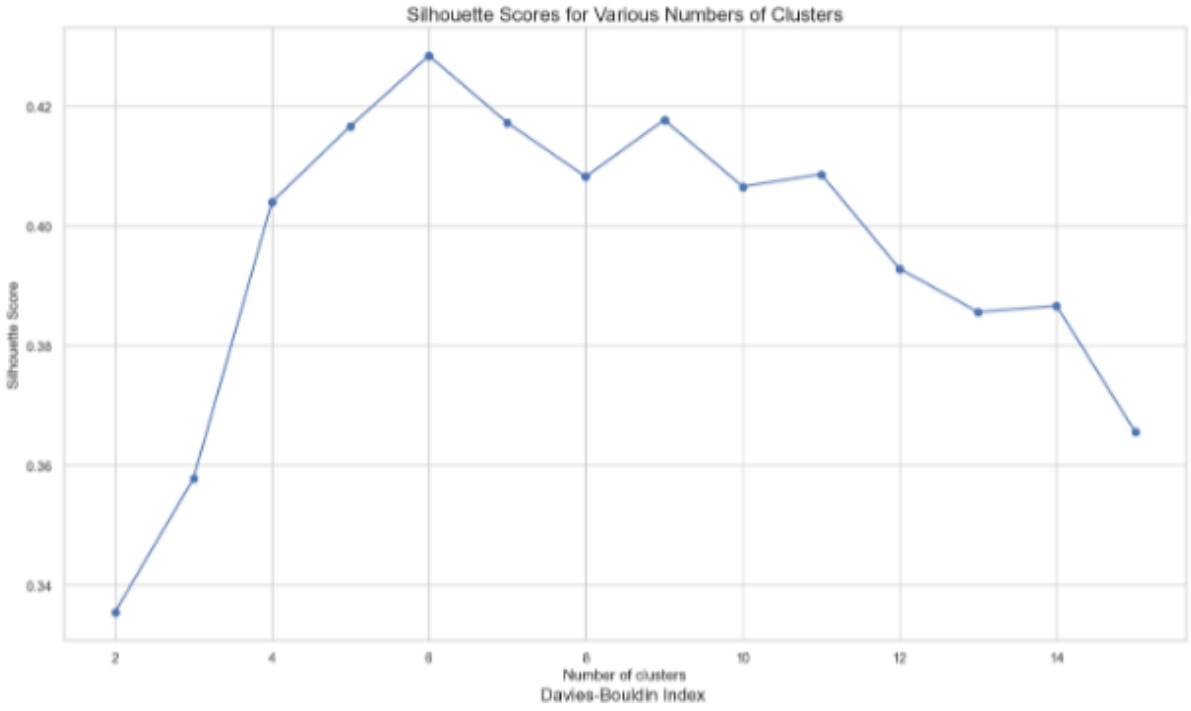

Age Distribution
Mean: 38.9
Count
Age


Age Distribution by Gender
Gender
Count
Age

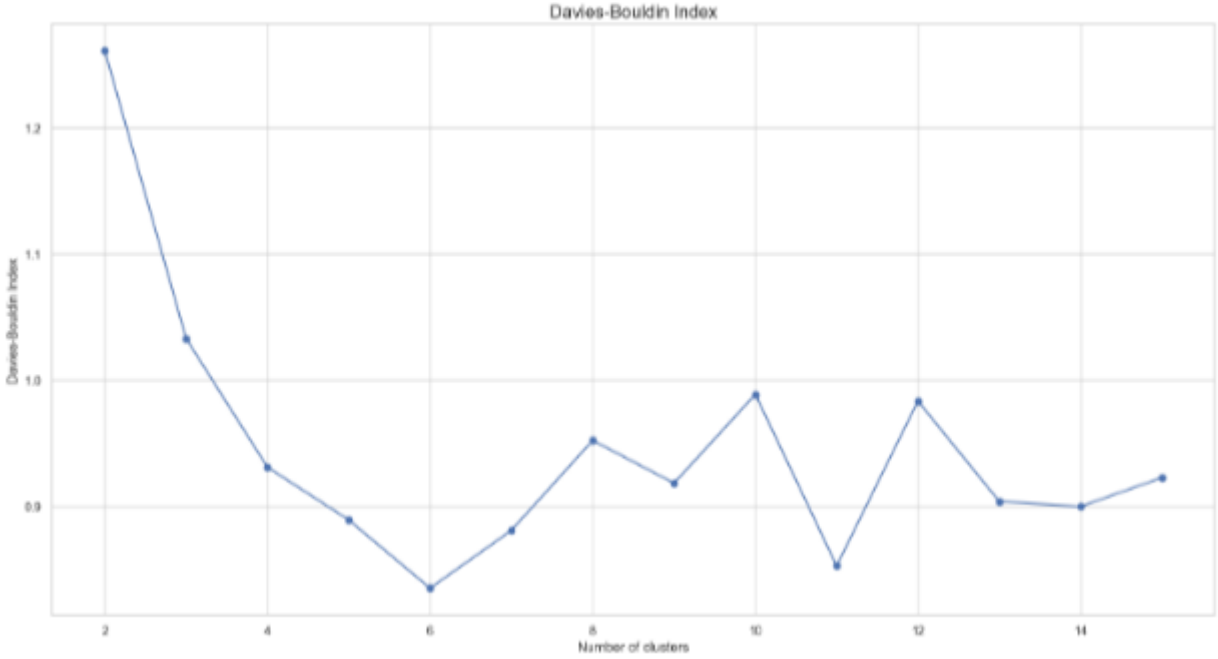

Comprehensive Annual Income Analysis
Annual Income Distribution
Annual Income vs Age Groups
Annual Income vs Gender
Count
Annual Income (k$)


Annual Income vs Age Groups
ageGroups
18-31
31-44
44-57
57-70
Count
0
5
10
15
20
25
30
35
20
40
60
80
100
120
140
Annual Income (k$)

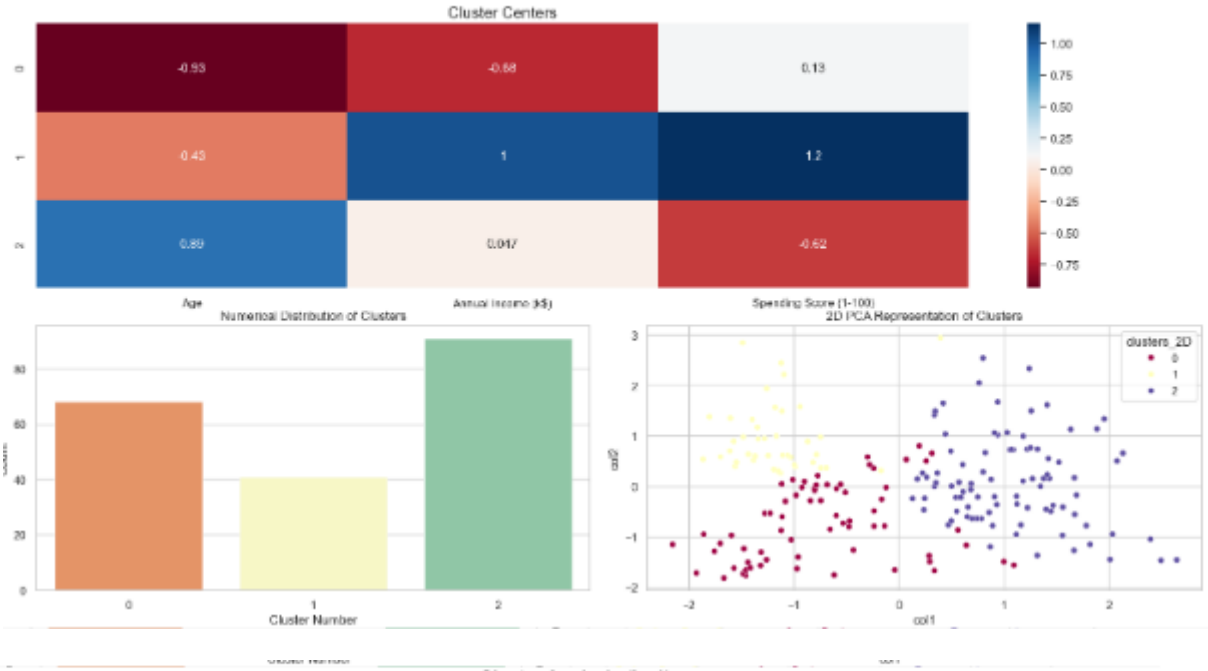

Annual Income vs Gender
Gender
Male
Female
Count
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20
40
60
80
100
120
140
Annual Income (k$)



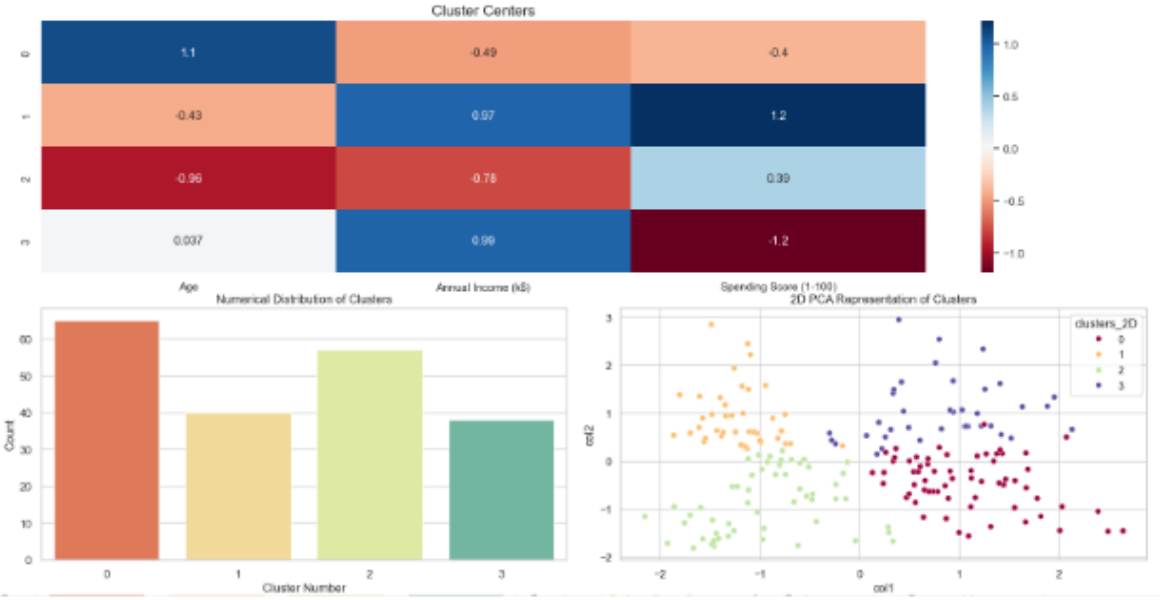
Davies-Bouldin Index



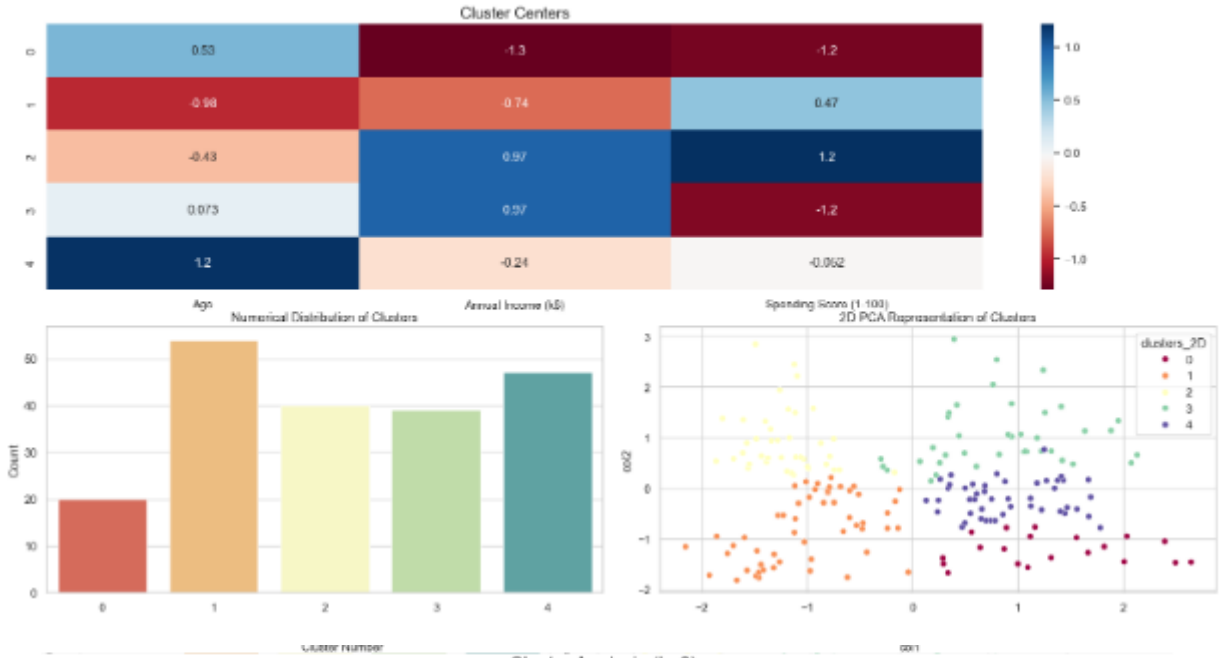
Cluster Analysis (k=3)



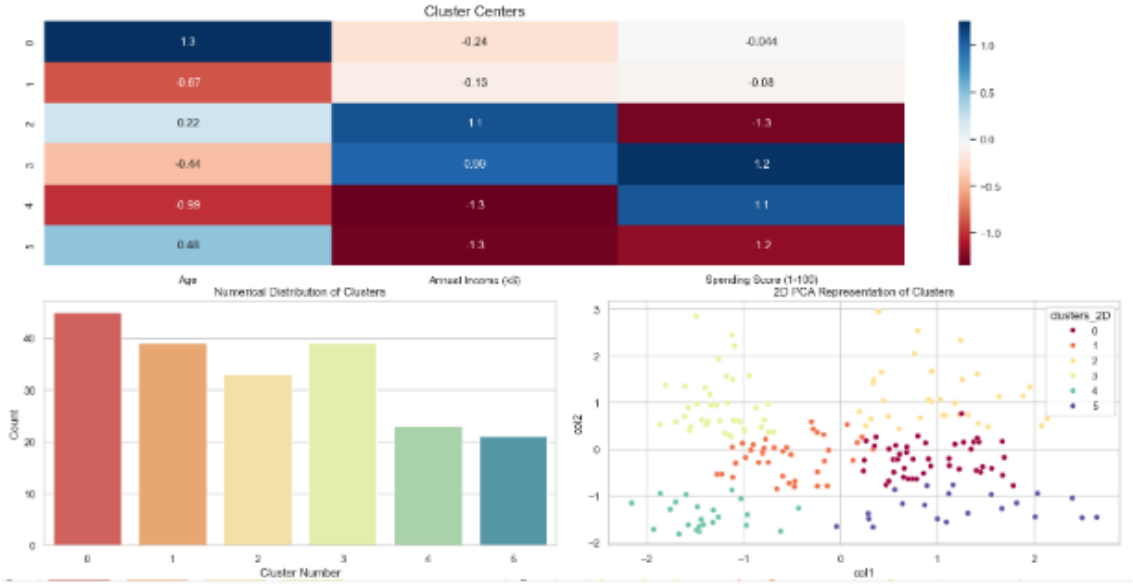
Cluster Analysis (k=4)



Cluster Analysis (k=5)



Cluster Analysis (k=6)



Customer Persona	Age	Income	Spending Score (1-100)
Leila	25	25k$ (LOW)	78 (HIGH)
Asma	56	54k$ (AVG)	49 (AVG)
Mohamed	42	89k$ (HIGH)	17 (LOW)
Dounia	27	57k$ (AVG)	48 (AVG)
Hakim	33	87k$ (HIGH)	82 (HIGH)
Ikram	46	26k$ (LOW)	19 (LOW)