


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Akli Mohand Oulhadj de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master 2

en Informatique

Spécialité : *GSI*

Thème

La préservation de la vie privée pour clustering

Encadré par

— Boudjelaba Hakim

Réalisé par

— Boumesli Bouchra

— Madani Fatima Zohra

2021/2022

Remerciements

On remercie Dieu LE TOUT PUISSANT de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tout d'abord, ce travail ne serait pas riche et n'aurait pas pu avoir le jour sans l'aide et l'encaderement de : Mr. BOUDJELABA , on le remercie pour la qualité de son encaderement exceptionnel, pour sa patience et sa disponibilité durant notre préparation de ce mémoire.

Aussi l'ensemble des membres de jury qui nous ont fait l'honneur de bien vouloir évaluer ce travail.

On tient à remercier tous les enseignants et les membres du département informatique pour leur sérieux et efforts durant toutes ces années.

On remercie également nos parents pour leur pationce et soutien moral tout au long de ce travail sans oublier de remercier nos frères et soeurs.

Dédicaces

Je remercie dieu de m'avoir donner la patience et le courage durant ces longues années d'études pour finir ce travail.

Je veux dédier ma thèse à ma famille, un sentiment particulier de gratitude à mes parents dont les mots d'encouragement tout au long de mes études et toute ma vie.

Mes frères qui ne m'ont jamais quitté et leurs soutien. Ma grand-mère qui nous a quitté récemment elle a toujours voulu que je sois la meilleur.

Je dédie également cette thèse pour tous ceux qui posaient toujours des questions sur moi.

Une dédicace spéciale pour tous mes amis qui voulaient mon succès.

Madani Fatima Zohra

Dédicaces

Je dédie ce travail,

A mes chers parents, qui ont toujours été à mes côtés et ont toujours cru en moi. Pour leur amour, et leur soutien permanent tout au long de mes années d'études. Que Dieu vous protège,

A mes très chers frères Messaoud et Oussama et mes soeurs Khadidja,Chahra et Maria.

A ma famille,mes proches et à ceux qui me donnent que l'amour et de la vivacité.

A tous mes amis qui m'ont toujours encoragé,et à qui je souhaite plus de succès.

A tous ceux que j'aime.

BOUMESLI Bouchra.

Résumé

La vie privée et la publication des données sont souvent considérées comme une contradiction. Le clustering est une méthode d'apprentissage non supervisée qui combine des éléments d'entrée similaires en groupes, utilisés dans de nombreux domaines allant de l'analyse commerciale aux soins médicaux. Dans bon nombre de ces applications, des informations sensibles sont collectées et qui en principe ne devraient pas être divulguées.

Dans notre travail, nous abordons le problème de l'anonymat des données appliqué au clustering de données, en d'autres termes, comment protéger un ensemble de données contre la ré-identification des individus tout en gardant la qualité des données pour le clusterning.

Notre méthode proposée consiste à perturbé le dataset original en utilisant la technique CTGAN qui permet de générer des données synthétique pour remplacer les données d'origine.

Mots clés : Clustering, L'anonymat, vie privée, ré-identification, protection des données personnelles, micro-données, dataset perturbé ...

Abstract

Privacy and Data publishing are often seen as contradictory.clustering is an unsupervised learning method that combines elements of similar groups, used in many fields ranging from business analysis to health care. In many of these applications, sensitive information are collected and which in principle should not be disclosed.

In our work, we address the problem of data anonymity applied in data clustering, in other words, how to protect a set of data against the re-identification of individuals while maintaining the quality of the data for the clustering.

Our proposed method consists in pertubed the dataset Original using the CTGAN technique which allows to generate synthetic data to replace the original data.

Key words : Clustering, Anonymity, privacy, re-identification, personal data protection,

micro-data, perturbed dataset ...

ملخص

غالبا ما ينظر إلى الخصوصية ونشر البيانات على أنهما متناقضان. التجميع هو طريقة تعلم غير خاضعة للرقابة تجمع بين عناصر الادخال المتشابهة في مجموعات ، وتستخدم في العديد من المجالات مثل تحليل الأعمال و الرعاية الصحية. في العديد من هذه التطبيقات ، يتم جمع المعلومات الحساسة والتي من حيث المبدأ لا ينبغي الكشف عنها .

في عملنا، نعالج مشكلة إخفاء هوية البيانات المطبقة في تجميع البيانات، بعبارات أخرى، كيفية حماية مجموعة من البيانات من إعادة تحديد هوية الأفراد مع الحفاظ على جودة البيانات للتجميع

تتمثل طريقتنا المقترحة في تشويه مجموعة البيانات الاصلية باستخدام تقنية CTGAN التي تسمح بإنشاء بيانات اصطناعية لتحل محل البيانات الأصلية.

الكلمات الرئيسية : التجميع ، عدم الكشف عن هويته ، الخصوصية ، إعادة تحديد الهوية ، حماية البيانات الشخصية ، البيانات الجزئية ، مجموعة البيانات المضطربة ...

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ICA	Independent Component Analysis
SVD	Singular value Dicomposition
TDP	Translation Data Perturbation
RDP	Rotation Data Perturbation
SDP	Scaling Data Perturbation
SMC	Secure Multiparty Computation
PCA	Principale Analysis Component
GAN	Generative Adversarial Network

Introduction générale

La protection des données peut être définie comme la protection des personnes, d'institutions et d'organisations réels qui doivent être protégés légalement et éthiquement pendant le cycle de vie des données (Collecte, traitement et analyse des données, diffusion, partage, conservation et réutilisation des données). Au cours de ce processus, les données peuvent être traitées, partagées, et peuvent aussi être transférées avec les personnes qui ont des droits d'accès sur ces données avec un niveau d'accès transparent et gérable, c'est là des exigences importantes pour la confidentialité des données. La vie privée, en revanche, n'a pas de définition précise, qui peut être spécifique à une application.

Les processus de collecte et de traitement des données doivent prendre des précautions en matière de confidentialité pour prévenir les violations de données et doivent être fiables et sont légalement tenus d'utiliser des méthodes appropriées pour stocker et utiliser les données collectées via des applications numériques et de les partager de manière anonyme si nécessaire.

En effet le clustering ou bien la classification non supervisée est une méthode d'apprentissage automatique qui consiste à regrouper des points de données par similarité ou par distance. Le clustering est largement utilisé dans le nombreux de domaines tel que la finance, la médecine et les réseaux sociaux.

La préservation de la vie privée des personnes lorsque des données sont partagées pour le clustering est un problème complexe. Le défi est de savoir comment protéger les valeurs de données sous-jacentes soumises au clustering sans affecter la similarité (l'utilité des données) entre les objets analysés.

Avec la croissance rapide des bases de données, des réseaux et des technologies informatiques, une grande quantité de données personnelles peut être intégrée et analysée numériquement, ce qui entraîne une utilisation accrue d'outils d'exploration de données pour déduire des tendances et des modèles. Cette évolution a suscité des préoccupations universelles concernant la protection de la vie privée des individus. Nous intéressons dans notre travail aux solutions d'anonymisation qui réduisent le risque de ré-identification. Le processus d'anonymisation doit être indépendant de l'utilisation des données. Celles-ci doivent être au même format que les données point de départ tout en conservant suffisamment d'informations pour permettre à l'utilisateur d'effectuer diverses analyses.

Le risque de réidentification se produit lorsque des informations d'identification personnelle peuvent être découvertes dans des données effacées ou "anonymisées".

Le but de ce travail est de proposer une nouvelle approche d'annonymisation des données pour le clustering en utilisant une approche d'anonymisation de donnée pour se protèger contre le risque de réidentification. Pour la réalisation de notre travail, nous allons suivre le plan suivant :

- **Premier chapitre** est consacré aux notions de bases de l'nonymat et de la vie privée.
- **Deuxiéme chapitre** présente une vue génerale sur le clustering ainsi que les methodes d'anonymats appliquées au clustering .
- **Troisième et dernier chapitre** nous allons présenter notre approche ainsi que les algorithmes de clustering choisi pour la validation de notre approche et les simulations réalisées pour tester l'efficacité de notre proposition.

Et nous terminerons par une conclusion générale et quelques perspectives.

Chapitre 1

Vie privée et anonymat

1.1 Introduction

Nous vivants dans un univers où rien ne peut être caché, et les moyens technologiques pour épier nos vies privées sont en nombre gigantesque, avec le temps on est devenu accros aux réseaux sociaux et aux nouvelles technologies de communication, donc nous devons assurer nous même notre sécurité en prenant beaucoup de réserve le terme anonymat qui n'a plus de sens à l'air du numérique.

Parfois, les gens acceptent volontiers de partager de l'information sur eux-mêmes comme partie normale des transactions Internet et de socialisation. Très souvent, les gens ignorent complètement qu'ils partagent de l'information, ce à quoi elle peut servir, ou la portée et le volume de l'information qui est partagée à leur sujet.

1.2 Définition de la vie privée

La vie privée signifie la préservation des informations ou des activités privées des utilisateurs et l'impossibilité de les obtenir, telle que l'identité de la personne, dossiers médicaux, conversations privées (mails, sms, etc.), photos et vidéos personnelles, le respect de la confidentialité des utilisateurs.

La vie privée concerne le contenu. Il s'agit de se soustraire au regard du public et de préserver la confidentialité. Exemple si vous envoyez un courrier électronique crypté à un

ami afin que seuls deux d'entre vous puissent l'ouvrir, il s'agit d'un message privé. Il n'est pas public [1].

1.3 Définition de l'anonymat

L'anonymat signifie la dissimulation ou l'absence d'identité (nom de personne) bien que son activité soient visibles par tout le monde. Par exemple une personne peut se connecter à un service d'anonymat comme ' Freenet' pour publier un message politique ou de partager des fichiers et de discuter sur des forums sous un nom d'utilisateur anonyme, dans ce cas, la personne peut publier un message public tout en gardant son identité anonyme [1].

1.4 Relation entre vie prive et anonymat

La vie privée et l'anonymat sont deux concepts différents mais liés. Ils sont tous deux de plus en plus nécessaires à mesure que nous sommes de plus en plus mis sur écoute et suivis, légalement ou non, et il est aussi important de comprendre pourquoi ils font partie intégrante de nos libertés civiques, pourquoi ils ne sont pas seulement bénéfiques pour l'individu, mais absolument essentiels à une société libre.

La protection de la vie privée est la capacité de garder certaines choses pour soi, indépendamment de leur impact sur la société. La recherche montre qu'elle va au-delà d'un désir et qu'elle est un besoin profond – dans toutes les sociétés à travers l'histoire, les gens ont créé des espaces privés pour eux-mêmes. Même dans les régimes les plus oppressifs, les gens ont trouvé un moyen de faire quelque chose, quelque chose de petit, en dehors des regards indiscrets. C'est assez révélateur. La vie privée est donc un concept qui décrit les activités que vous réservez entièrement à vous-même ou à un groupe limité de personnes.

Par contre l'anonymat, c'est quand vous voulez que les gens voient ce que vous faites, mais pas que c'est vous qui le faites. L'anonymat est donc une sorte de préservation de vie privée car nous souhaitons garder une information rien qu'à nous-mêmes qui est notre identité, nous pouvons donc dire que l'anonymat est inclus dans la vie privée.

L'exemple typique serait si vous voulez dénoncer l'abus de pouvoir ou d'autres formes de criminalité dans votre organisation sans risquer sa carrière et son statut social dans ce groupe, et c'est pourquoi nous avons généralement des lois solides qui protègent les sources de la presse libre. Vous pouvez également publier ces données anonymement en ligne via un VPN, le réseau d'anonymisation Freenet, ou les deux. C'est l'équivalent analogique de la lettre de dénonciation anonyme, qui a été considérée comme un régime de base dans nos freins et contrepoids. Il est évident que ces concepts "*Vie privée et Anonymat*" sont bénéfiques pour l'individu. Mais plus important encore, il est dans l'intérêt général de la société que chaque individu bénéficie de ses avantages. Il n'y a pas seulement un avantage individuel, mais un avantage collectif [2].

1.5 Les attaques de la vie privée

Toutes les menaces relatives à la vie privée tournent autour de l'utilisation non autorisée et/ou malveillante des données collectées (d'une manière légale ou illégale). Ces menaces peuvent être :
Divulgation des données personnelles (Atteinte à la réputation et à l'intimité) - Avec l'émergence des réseaux sociaux, les services de partages des photos et des vidéos, trouver et accéder à des informations personnelles est devenu une opération très simple. Des informations comme la date de naissance, l'emploi, la situation familiale, les préférences musicales, les informations comportementales, etc. qui sont considérées par la majorité d'entre nous comme triviales et inoffensives peuvent être utilisées contre nous. En accédant à ces informations, les risques de préjudice, d'inégalité, de discrimination et de perte d'autonomie apparaissent facilement, les exemples dans ce cadre ne manquent pas.

Vol d'identité

Est un crime dans lequel un pirate va acquérir les données personnelles : des noms, des adresses et des numéros de téléphone, des numéros d'assurance sociale, des numéros de permis de conduire, des renseignements sur des cartes de crédit et des cartes bancaires[3].

Le vol d'identité consiste à utiliser les renseignements personnels d'un individu à son insu. Ce crime est souvent perpétré simultanément à d'autres crimes tels la fraude, la contrefaçon et le vol. Les cibles attrayantes pour un usurpateur d'identité comprennent

le numéro d'assurance sociale (NAS), le numéro de permis de conduire, les cartes de crédit, les cartes de débit, les chéquiers, les cartes de téléphone, les mots de passe, les NIP (numéros d'identification personnelle), etc.

En règle générale, ces personnes tentent de faire des transactions, d'obtenir du comptant, de la marchandise ou des services avant que leur identité ne puisse être détectée ou que le propriétaire véritable des renseignements s'en rende compte [4].

Le Profilage

Chaque fois que l'utilisateur visite un site Web, quelqu'un, quelque part suit son activité en ligne. Ce profilage permet de recueillir des renseignements détaillés sur l'internaute et cette pratique est très utilisée sur de nombreux sites, souvent à l'insu du visiteur ou sans son consentement, et cela présente un risque d'atteinte à la vie privée du fait qu'il permet d'analyser avec précision le comportement des consommateurs. Les informations utilisées dans le profilage contiennent : des données d'identificateurs tels-que : les adresses IP, les numéros d'identification des navigateurs web et les systèmes d'exploitations, etc. Et des données concernant les activités et le comportement des utilisateurs sur Internet, comme : les requêtes sur les moteurs de recherche, les sites visités, les relations et les communications sur les réseaux sociaux, les produits achetés sur Internet, etc.

Le profilage est une technique de surveillance ou d'exploitation des données qui permet d'établir différentes actions, mesures ou décisions touchant les personnes concernées dans le cadre de finalités diverses. Les techniques de profilage représentent un intérêt important pour l'économie ou pour les administrations publiques ; elles peuvent aussi avoir des effets bénéfiques pour les personnes concernées, par exemple dans le domaine de la santé. Cependant, elles génèrent également des conséquences négatives sur le respect des droits et des libertés fondamentales, notamment le droit à la vie privée et à la protection des données [5].

1.6 L'anonymisation de micro-données

Micro données désigne un enregistrement contenant des informations relatives à un individu spécifique (un citoyen ou une entreprise). Une publication de micro données vise

à publier des données brutes, c'est-à-dire un ensemble d'enregistrements de micro données.

On peut retrouver dans une micro-donnée quatre groupes distincts d'attributs : l'identifiant explicite, le quasi-identifiant, le groupe d'attributs sensibles et le groupe d'attributs non sensibles.

La solution la plus couramment utilisée par les organisations pour protéger les micodonnées serait de supprimer les attributs de type identifiant ou pseudonymisation avant de publier les tables. Cependant, cette méthode est insuffisante pour préserver la vie privée des individus concernés lorsqu'ils sont capables de combiner différentes bases de données pour "ré-identifier" les individus. Plusieurs situations particulières l'ont démontré. Citons, à titre d'exemple, Sweeney l'illustre dans son étude plusieurs exemples dont En Massachusetts, la Groupe Insurance Commission est responsable de la mise en place d'une assurance maladie pour les employés du gouvernement. il avaient une liste d'inscriptions des Electeurs de Cambridge et Massachusetts contenant les informations. Le nom, la résidence, le code postal, la date de naissance et le sexe de chaque électeur ont été inclus dans les données, comme le montre le cercle le plus à droite de la figure 1. Ces données peuvent être liées aux dossiers médicaux à l'aide des codes postaux, des dates de naissance et du sexe, permettant de diagnostiquer des procédures et des prescriptions à lier à des individus spécifiques.

Par exemple, à l'époque, William Wells était gouverneur du Massachusetts et ses informations médicales étaient incluses dans la base de données GIC. Le gouverneur Wells était un résident de Cambridge, Massachusetts. Selon la liste des électeurs de Cambridge, il n'y avait que six personnes avec sa date de naissance, trois d'entre eux étaient des hommes et il était le seul dans son code postal à cinq chiffres [6].La fégure suivante présente lien pour ré-identifier les données 1.1 .

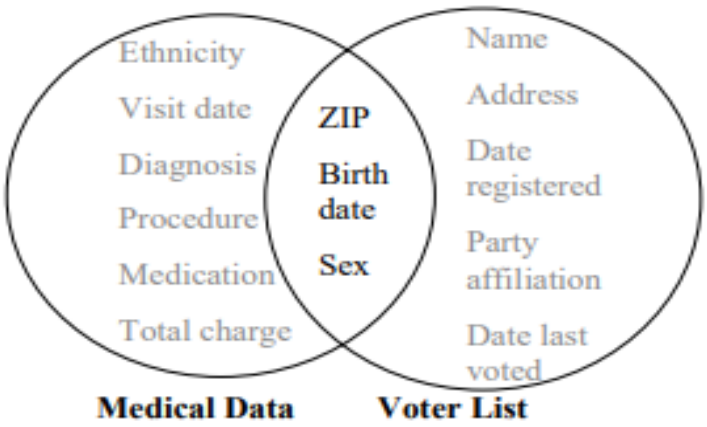


FIGURE 1.1 – Lien pour ré-identifier les données[6]

1.7 Les techniques de protection de la vie privée

De nombreux modèles de protection ont été proposés dans la littérature de la Publication de Données Respectueuse de la Vie Privée. Dans cette partie, nous ne citons que quelques exemples.

1.7.1 k-anonymat

Le concept de k-anonymat a été introduit pour la première fois en 1998 par Latina Sweeney pour remédier au risque de réidentification des données anonymisées via la combinsant de différentes bases de données et pour s'assurer que les deux catégories de données ne peuvent pas être connectées l'une à l'autre [6].

k-anonymat repose sur l'idée qu'en combinant des enregistrements avec des attributs similaires pour que chaque enregistrement ne puisse pas être distingué d'au moins k-1 autres enregistrements et que ces k enregistrements forment une classe d'équivalence [6].

Les deux principes de base pour transformer des ensembles de données en tables k-anonymes sont la généralisation et la suppression :

La généralisation

Est la pratique consistant à substituer une valeur spécifique à une valeur plus générale. Par exemple, les ensembles de données qui incluent des codes postaux peuvent généraliser

des codes postaux spécifiques dans des counters ou des municipalités (c.-à-d. en changeant 01234.En 012**). Les âges peuvent être généralisés dans une tranche d'âge (c'est-à-dire regrouper "Âge : 35" en "Groupe d'âge : 30-39").

Suppression

La suppression consiste à supprimer entièrement la valeur d'un attribut d'un ensemble de données. Dans l'exemple de données sur l'âge, la suppression signifierait la suppression complète des informations sur l'âge de chaque cohorte. la suppression ne doit être utilisée que pour les données qui ne sont pas pertinentes pour l'objectif de la collecte de données. Ci-dessous (tables 1.1et 1.2) un exemple d'une table k-Anonyme avec k =3 et les quasi-identificateurs = code postal, âge , salaire et attribut diagnostique comme donnée sensible.

Code postale	**Age**	**Salaire**
1598	23	30k
1678	28	80k
1637	30	100k
1616	32	150k
1525	25	30k

TABLE 1.1 – table orignale.

Code postale	**Age**	**Salaire**
15**	[22-25]	30k
15**	[22-25]	30k
16**	>=28	100k
16**	>= 28	150k
16**	>= 28	80k

TABLE 1.2 – table 2-anonymat

Points faibles

On pourrait dire que le k-Anonymat nous offre la protection et la confidentialité, mais la vérité est que la méthode est vulnérable à de nombreuses attaques et elle a des limitations. Prenons un exemple :

- Perte d'informations.
- Attaque de correspondance non triée : le problème est que l'ordre des enregistrements est le même dans la table publiée et dans la table d'origine. La solution à ce problème consiste à randomiser l'ordre des enregistrements avant de publier une base de donnée k-anonymat.
- Attaque d'homogénéité : Cette attaque exploite la situation où toutes les valeurs d'une valeur sensible dans un ensemble de k enregistrements sont les mêmes. Dans ce cas, les valeurs de sensibilité de tous les k enregistrements peuvent être prédites même si les données ont été anonymisées.
- Attaque de connaissances de base : Dans ce type d'attaque, l'adversaire a une connaissance connue de l'individu et avec un raisonnement logique supplémentaire, les attributs sensibles de l'individu peuvent être divulgués. Dans cette attaque, l'adversaire peut utiliser une association entre un ou plusieurs attributs de quasi-identifiant avec l'attribut sensible ou la connaissance publique de la cible afin d'éliminer les valeurs possibles de l'attribut sensible.

1.7.2 *l*-diversité

l-diversité définie dans machnavajjhala et al en 2007, est une extension du modèle k-anonymat mais garanti aussi la diversité pour les valeurs sensibles dans le mécanisme d'anonymisation, elle permet de contrer des attaques par homogénéité et des attaques fondées sur des connaissances de base.

Une classe d'équivalence est dite à *l*-diversité s'il existe au moins l valeurs distinctes pour l'attribut sensible. En fait, si nous n'effectuons pas cette vérification, il est possible que la valeur sensible ne soit que *l*-diverse donc qu'on puisse retrouver la valeur sensible liée à un individu, bien que cet individu ait été anonymisé par la technique de k-anonymisation [7].

Une base de données est dite *l*-diverse si toutes ses classes d'équivalence a une *l*-diversité.

Code postale	**Age**	**Sexe**	**Diagnostique**
1300	23	F	Arythmie
1256	28	M	Diabète
1390	21	F	Hypertension
1200	25	M	Cancer
1289	26	F	VIH
1320	24	F	Athérosclérose
1234	28	F	rythme

TABLE 1.3 – Table originale.

Code	**Age**	**Sexe**	**Diagnostique**
13**	[20-24]	F	Arythmie
13**	[20-24]	F	athérosclérose
13**	[20-24]	F	Hypertension
12**	[25-29]	F	Rhume
12**	[25-29]	F	VIH
12**	[25-29]	M	diabète
12**	[25-29]	M	Cancer

TABLE 1.4 – Table anonymisé avec 2-diversité.

Pour la table 1.4 on a chaque classe d'équivalence de la table a au moins 2 (L) valeurs distinctes pour l'attribut sensible.

Point faible

- *l*-diversité n'assure pas la protection contre les attaque par similarité et les attaques par inférences et probabilistes.
- Attaque par similarité : les valeurs de l'attribut sensible dans une classe d'équivalence sont distinctes syntaxiquement différentes mais sémantiquement similaires.
- Attaque d'asymétrie : Lorsque la distribution globale est faussée, la diversité satisfaisante n'empêche pas la divulgation des attributs.
- Perte de précision.

1.7.3 t-proximité

La t-proximité est un affinement de la diversité l, car elle vise à créer une classe d'équivalence similaire à la distribution initiale des attributs dans un tableau. Cette technique est utile lorsqu'il est important de garder les données aussi proches que possible de l'original. En d'autres termes, il introduit le concept de distance entre ces deux distributions ; pour cela, une contrainte supplémentaire est ajoutée aux classes d'équivalence que non seulement chaque classe d'équivalence doit contenir au moins l valeurs distinctes, mais que chaque valeur Les deux doivent être représentés autant de fois que nécessaire pour refléter la distribution initiale de chaque attribut.

Une classe d'équivalence est dite à t-proximité si la distance entre la distribution d'un attribut sensible dans cette classe et la distribution de l'attribut dans la table entière n'est pas supérieure à un seuil t. Une table est dite à proximité t si toutes les classes d'équivalence ont une t-proximité[8]. le problème dans cette méthode est que si la taille et la variété des données augmentent, les chances de ré-identification du processus augmentent également [9].

Point faible

- t-proximité protège contre la divulgation d'attributs, mais ne traite pas de la divulgation d'identité.
- t-proximité traite de l'homogénéité et de la connaissance du contexte.
- elle n'est certainement pas parfaite.
- la relation entre la valeur t et le gain d'information n'est pas claire.
- t-proximité inclura la perte d'utilité des données.

1.7.4 δ-présence

Elle a été proposé par Nergiz, Atzori, et Clifton en 2007, Ce modèle forcé que la probabilité de présence d'un enregistrement soit dans un intervalle $\delta = (\delta \text{ min}, \delta \text{ max})$ prédéfini.

Pour faire une protection contre les attaques par « lien de tables ». Pour empêcher ce type d'attaque et donc éviter d'inférer la présence de tout enregistrement d'une table publiée dans une autre table publiée. Ce modèle bien qu'intéressant est difficile à mettre

en œuvre car il suppose que l'éditeur connaisse a priori la table de rapprochement que l'attaquant est susceptible d'utiliser [10].

1.8 Les techniques d'anonymat

Il existe différentes techniques d'anonymisation. Lors de l'anonymisation, vous pouvez décider d'en utiliser une ou plusieurs selon votre volume de données et vos besoins.

1.8.1 Généralisation

Cette technique consiste à créer une hiérarchie de généralisation qui permettra de transformer les attributs QI en attributs moins précises et plus générales, pour que doter pour chaque attribut du QI d'un arabe de généralisation, afin d'assurer que tous les enregistrements soient identiques à au moins k-1 autres [11]. Figure 1.2 Hiérarchie de gé-

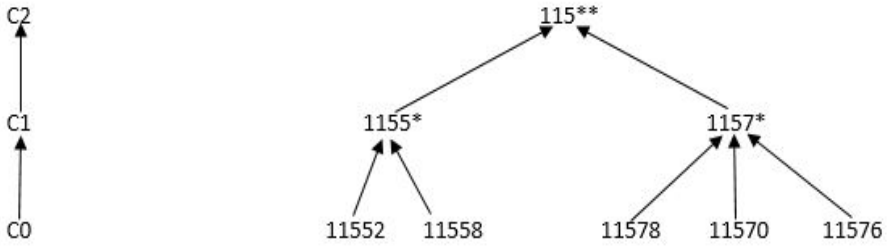


FIGURE 1.2 – Hiérarchie de généralisation de l'attribut code postal

néralisation pour l'attribut de code POSTAL. En bas, C0 représente le domaine d'origine (non généralisé) de l'attribut. La première généralisation, C1, regroupe les codes postaux dont les quatre premiers chiffres correspondent. La deuxième généralisation, C2, regroupe tous les codes postaux de l'ensemble de données. Une fois les hiérarchies de généralisation définies pour chaque attribut individuel, nous les combinons pour obtenir une généralisation d'enregistrement (c'est-à-dire que nous sélectionnons une généralisation pour chaque attribut).

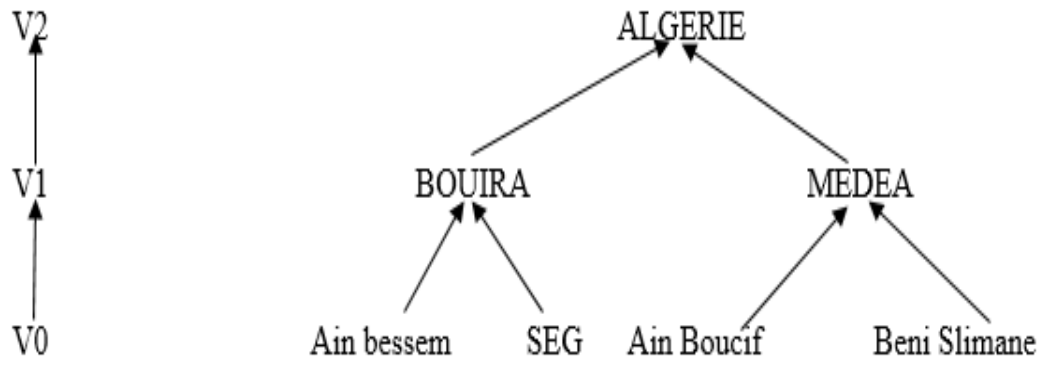


FIGURE 1.3 – Hiérarchie de généralisation de l'attribut ville.

1.8.2 La suppression

Suppression signifie supprimer des données de la table afin qu'elles ne soient pas libérées. Une suppression peut concerner de la suppression de tuple ou une attribuée ou suppression de quelques données de tuples (généralement remplacement par *). La suppression est utilisée en complément de la généralisation pour réduire la quantité de généralisation requise pour obtenir la propriété k-anonymat (exemple dans cas des tuples avec moins de k occurrences) [11]. La suppression permet de satisfaire les exigences de k-anonymat avec moins de généralisation.

1.8.3 Permutation

peut être appliquée aussi bien sur un attribut du QI que sur un attribut sensible, qu'il soit continu ou catégoriel est utile lorsqu'il est important de conserver la distribution exacte de chaque attribut dans l'ensemble de données. Les techniques de permutation modifient les valeurs de l'ensemble de données en les échangeant simplement d'un enregistrement à un autre. Un tel échange garantira que la plage et la distribution des valeurs resteront les mêmes, mais les corrélations entre les valeurs et les individus ne le seront pas. La permutation peut ne pas fournir une anonymisation en elle-même et doit toujours être combinée à la suppression d'attributs / quasi-identifiants évidents [12].

Le tableau 3 représente un extrait de données originales médicales. La table est constituée de huit attributs : sexe, âge, ville, diagnostique. Exemple, la permutation appliquée sur l'attribut sexe au sein du sous-ensemble constitué des tuples 1 et 5 donnerait la table

anonyme suivante :

Age	**Ville**	**Sexe**	**Diagnostique**
46	Bouira	F	Cirrhose
31	Bouira	M	Bronchite
68	Médéa	F	Cancer de sein
96	Bejaia	F	Hépatite
17	Bouira	F	diabète
46	Médéa	F	Bronchite
42	Bejaia	M	Grippe

TABLE 1.5 – Les données originales.

Age	**Ville**	**Sexe**	**Diagnostique**
46	Bouira	F	Cirrhose
46	Bouira	M	Bronchite
68	Médéa	F	Cancer de sein
96	Bejaia	F	Hépatite
17	Bouira	F	diabète
31	Médéa	M	Bronchite
42	Bejaia	M	Grippe

TABLE 1.6 – les données anonymisé par swapping.

1.8.4 Micro-agrégation

Cette technique est fondée sur une classification des enregistrements correspondent en plusieurs group dont l'effectif est au moins k individus appelés micro-agrégats pour satisfaire la confidentialité des données les valeurs originales sont remplacée par une mesure centrale du micro- agrégat auquel elles appartiennent. Cette technique a d'abord été définie pour les données continues et étendue aux données catégorielles [13].
Quel que soit le type de données, la micro-agrégation peut être définie opérationnellement en fonction des deux étapes suivantes [13] :

Partitionnement

L'ensemble des enregistrements d'origine est partitionné en plusieurs clusters de telle sorte que les enregistrements d'un même cluster soient similaires les uns aux autres et que le nombre d'enregistrements dans chaque cluster soit d'au moins k. Cette étape doit mettre en place des groupes aussi homogènes que possible.

Agrégation

Un opérateur d'agrégation (par exemple, la moyenne pour les données continues ou la médiane pour les données catégorielles) est calculé pour chaque cluster et est utilisé pour remplace.les enregistrements d'origine. En d'autres termes, chaque enregistrement d'un cluster est remplacé par la valeur agrégée calculée pour le groupe auquel il appartient.

A titre d'exemple, nous allons appliquer la micro-agrégation à l'attribut 'sexe' dans le Tableau 3. La première étape consiste à diviser les enregistrements en groupes homogènes selon l'attribut sexe afin de satisfaire le 3-anonymat. Ensuite, nous allons remplacer la

Age	**Ville**	**Sexe**	**Diagnostique**
46	Bouira	F	Cirrhose
68	Médéa	F	Bronchite
96	Bejaia	F	Cancer du sein
46	Médéa	F	Hépatite
31	Bouira	M	Bronchite
17	Bouira	M	diabète
42	Bejaia	M	Grippe

TABLE 1.7 – Etape de partition de micro-agrégation

valeur de l'attribut 'age' de chaque enregistrement par la moyenne du groupe comme le montre le Tableau 1.8 Etape d'agrégation de la technique de micro-agrégation.

Age	**Ville**	**Sexe**	**Diagnostique**
64	Bouira	F	Cirrhose
64	Médéa	F	Bronchite
64	Bejaia	F	Cancer du sein
64	Médéa	F	Hépatite
30	Bouira	M	Bronchite
30	Bouira	M	diabète
30	Bejaia	M	Grippe

TABLE 1.8 – Étape de partition de micro-agrégation

1.9 Pourquoi anonymat sur internet ?

Pour la grande majorité des gens, la vie privée en ligne et la vie réelle sont différentes, nous autorisons certaines choses que nous ne pouvons même pas imaginer dans la vie réelle, mais nous penserons que les gens ne sont pas vraiment intéressés, peut-être par manque de connaissances. Cependant, la confidentialité en ligne est très importante [14]. Voici quelques raisons pour lesquelles vous devriez vous en soucier :

1.9.1 Vie privée – Données personnelles

Les données personnelles sont l'essence de l'économie numérique, les données collectées auprès des internautes enrichissent les entreprises du numérique, malheureusement nous avons vu de nombreux scandales et les entreprises ne peuvent garantir à 100% la sécurité des données des utilisateurs. Les gouvernements et les entreprises dressent le profil des internautes, et si ces informations finissent entre de mauvaises mains, il serait théoriquement possible de les manipuler pour changer leur façon de penser ou même voter, et on pense surtout que la Russie s'est ingérée lors de l'élection présidentielle. En 2016, cela a sans aucun doute joué un rôle important dans l'orientation de l'opinion publique des internautes [14].

1.9.2 Protection de l'identité

Internet est disponible dans tous les pays du monde, néanmoins il ne peut être utilisé de la même manière partout, Parfois, on ne veut tout simplement pas que quelqu'un sache qui on est vraiment. Même si on n'est pas impliqué dans quelque chose d'illégal ou de douteux. Il y a un niveau de sécurité sociale qui vient avec l'anonymat [15].

1.9.3 Censure

Certains pays censurent régulièrement l'accès à certains sites web, les réseaux sociaux comme Facebook, Twitter, Instagram et autres par exemple sont bloqués en chine. L'anonymat peut-être aussi utilisé lors d'une utilisation en entreprise, dans laquelle l'entreprise a mis en place une politique d'accès restrictive à Internet [16].

1.9.4 Couvrir des actions illicites ou réprimées

Les lanceurs d'alertes et leurs actes, qui consistent à divulguer des informations qu'ils jugent menaçant pour l'intérêt général ou public, sont clairement réprimées et sévèrement sanctionnés par ceux qui protègent ces informations surtout s'il s'agit d'un état, on pense notamment à l'affaire Snowden. Dans ce cas, la diffusion de ces informations aux médias nécessite souvent un canal de communication complètement anonyme pour échapper à une identification [16].

1.10 Conclusion

La facilité et la transparence du flux d'informations sur Internet ont accru les préoccupations concernant la vie privée et la sécurité des données, afin de réduire le risque de divulgation de la vie privée, des technologies d'anonymisation ont été proposées et sont utilisées pour remédier à ce problème.

Dans ce premier chapitre nous avons présenté quelques méthodes qui permettent d'assurer en quelques sortes la vie privée et l'anonymat. Dans le prochain chapitre nous allons parler des méthodes d'anonymat appliquées dans le cas du clustering qui essaient de protéger les données contre la ré-identification tout en gardant l'utilité des données.

Chapitre 2

Clustering et anonymat

2.1 Introduction

Avant que l'intelligence artificielle ne devient capable de détecter des similarités entre individus, ce sont bien des intelligences humaines qui ont implémenté les algorithmes de clustering.

Le clustering est une technique d'exploration de données bien connue pour la reconnaissance de formes et la recherche d'informations. L'ensemble de données à agréger peut contenir des données catégorielles ou numériques.

2.2 Clustering

Le clustering [17]Est le processus de regroupement d'un ensemble d'objets de données non classé en plusieurs groupes ou clusters afin que les objets d'un cluster présentent une similitude élevée, mais soient très différents des objets d'autres clusters. Les dissemblances et les similitudes sont évaluées en fonction des valeurs d'attribut décrivant les objets et impliquent souvent des mesures de distance. C'est une méthode d'apprentissage automatique non supervisée.

De plus, les objets dont la classe est inconnue sont généralement disponibles en grand nombre. C'est pourquoi, les approches non supervisées sont massivement utilisées pour traiter des données de manière automatique 3.2.

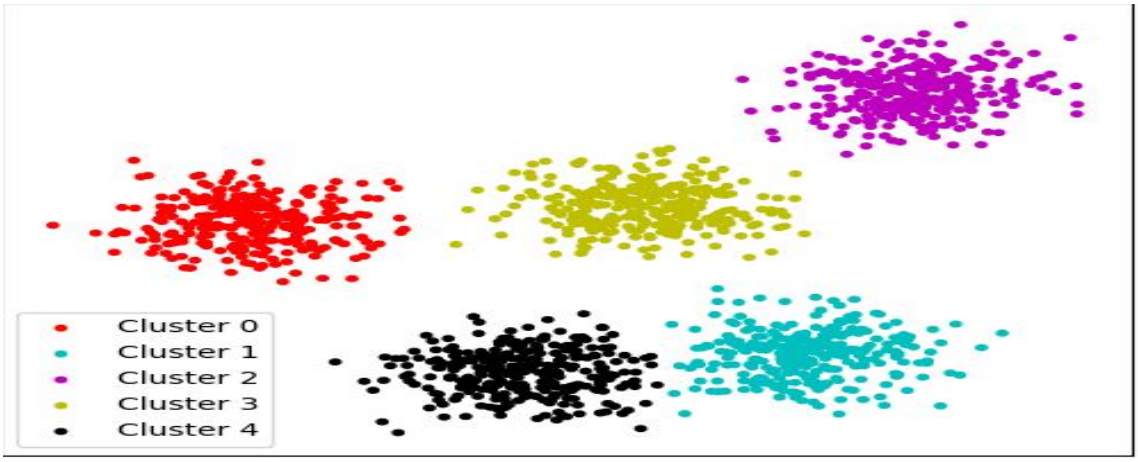


FIGURE 2.1 – Le clustering [18].

2.3 Principales étapes de clustering

Le processus de clustering se divise en trois étapes principales 2.3 :

2.3.1 Préparation des données

L'étape de préparation (prétraitement) des données consiste à sélectionner et/ou pondérer ces variables, voire à créer de nouvelles variables, afin de mieux discriminer entre eux les objets à traiter. L'existance des variables influentes ne sont pas nécessairement toutes pertinentes : certaines peuvent être Redondants et d'autres non-pertinentes pour la tâche ciblée. Cette question des variables a été largement étudié en classification supervisée, mais c'est encore un problème d'actualité dans les méthodes non supervisée [19].

2.3.2 Le choix de l'algorithme

Dans cette étape on va choisir l'algorithme de **clustering** le plus adapté pour le partitionnement de nos données. Ce choix est très lié au type de données à traiter. Le choix de l'algorithme de clustering doit donner lieu à une analyse globale du problème : quelle est la nature (qualitative et quantitative) des données ? Quelle est la nature des clusters attendus (nombre, forme, densité, etc.) ? L'algorithme doit alors être choisi de manière à ce que ses caractéristiques répondent convenablement à ces deux dernières questions. Les critères de décision peuvent être : la quantité de données à traiter, la nature de ces données, la forme des clusters souhaités ou encore le type de schéma attendu (pseudo-partition, partition stricte, dendrogramme, etc.) [19].

2.3.3 L'exploitation des clusters

La tentation est grande, pour un non-spécialiste, de considérer comme "acquis" le résultat d'un processus de clustering. Autrement dit, les clusters obtenus ne sont généralement ni remis en cause ni évalués en terme de disposition relative, dispersion, orientation, séparation, densité ou stabilité. Pourtant, il est sans aucun doute utile de distinguer les classes pertinentes obtenues, des autres. De même, cette étape d'analyse permet d'envisager le recours à une autre approche de clustering plus adaptée. Deux situations sont possibles : soit la tâche de clustering s'inscrit dans un traitement global d'apprentissage, soit les clusters générés par clustering constituent un résultat final [19].

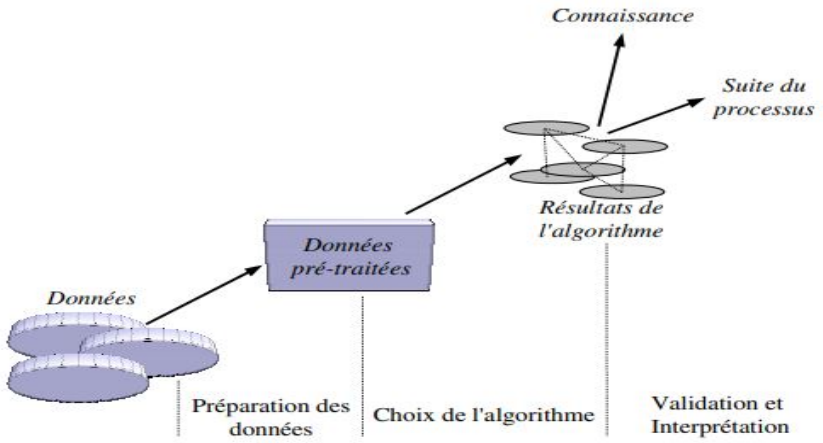


FIGURE 2.2 – les différentes étapes de clustering [19]

2.4 Les methodes de clustering

Les méthodes de clustering peuvent être classées en quatre catégories majeures.

2.4.1 Les methodes hiérarchique

Dans les algorithmes de clustering hiérarchique [20], les objets de données sont partitionnés en niveaux dans un format hiérarchique. Les clusters sont formés de manière itérative dans une approche descendante ou ascendante pour générer un dendrogramme représentant la structure hiérarchique des clusters formulés. Cette méthode de clustering permet d'explorer des données à différents niveaux de granularité. L'approche ascendante est appelée la méthode agglomération, tandis que l'approche descendante est la méthode

FIGURE 2.3 – Les methodes de clustering

de division. Dans la méthode agglomération, les clusters sont construits à partir d'objets uniques qui sont fusionnés de manière itérative de manière appropriée en clusters plus grands qui forment les différents niveaux de la hiérarchie jusqu'à ce que l'objet entier forme un seul cluster ou que le critère d'arrêt soit satisfait. L'inverse est le cas dans une méthode de division. Le cluster contenant tous les objets est décomposé de manière itérative de manière appropriée jusqu'à ce que chaque objet forme un seul cluster ou que le critère d'arrêt soit satisfait. La fusion ou la division est effectuée en fonction de la similitude ou de la dissimilarité des éléments du cluster. Figure 1. Illustre une représentation de dendrogramme pour la méthode de clustering hiérarchique 2.4.

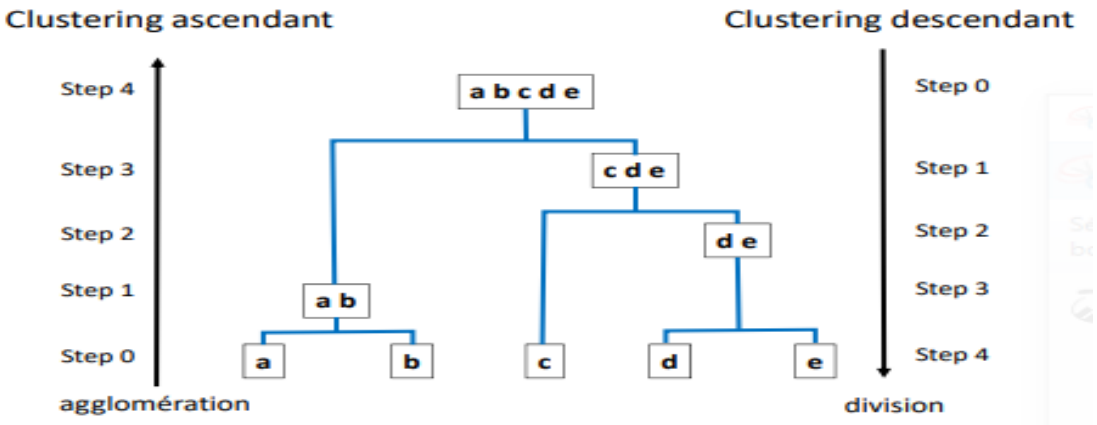


FIGURE 2.4 – Les methodes de clustering hiérarchique

2.4.2 Méthodes de partition

Ce type d'approche est basée sur l'idée de diviser un ensemble de données en K clusters non imbriqués, de sorte que chaque objet de données n'appartienne qu'à un seul cluster avec le centre le plus proche. Grâce à des itérations, les points de données sont réinitialisés avec un centre mis à jour pour former une meilleure partition, par réaffectant des objets autour des centres en mouvement, de telle sorte que tout point d'un cluster devient plus proche que tout les points extérieurs. Les clusters sont représentés par leur centroïdes qui correspondent à la moyenne de l'ensemble les objets contenus dans le cluster [21].

2.4.3 Méthodes basée sur la densité

Les méthodes de partitionnement et hiérarchiques sont conçues pour trouver des clusters de forme sphérique. Ils ont de la difficulté à trouver des amas de forme arbitraire tels que la forme en "S" et les amas ovales. Compte tenu de ces données, ils identifieraient probablement de manière inexacte les régions convexes, où le bruit ou les valeurs aberrantes sont inclus dans les clusters. Pour trouver des clusters de forme arbitraire, d'autres méthodes de clustering ont été développées sur la base de la notion de densité. Leur idée générale est de continuer à faire croître un cluster donné tant que la densité (nombre d'objets ou de points de données) dans le "voisinage" dépasse un certain seuil. Par exemple, pour chaque point de données d'un cluster donné, le voisinage d'un rayon donné doit contenir au moins un nombre minimum de points. Une telle méthode peut être utilisée pour filtrer le bruit ou les valeurs aberrantes et découvrir des amas de forme arbitraire [22].

2.4.4 Méthodes basée sur grille

Les méthodes basées sur une grille utilise des structures de données multi-résolution ,ou l'espace d'objet est quantifié en un nombre fini de cellules qui forment une structure de grille. Toutes les opérations de regroupement sont formées sur la structure de grille (c'est-à-dire sur l'espace quantifié). Le principal avantage de cette approche est son temps de traitement rapide, qui est typiquement indépendant du nombre d'objets de données et ne dépend que du nombre de cellules dans chaque dimension de l'espace quantifié. L'utilisation de grilles est souvent une approche efficace pour de nombreux problèmes

d'exploration de données spatiales, y compris le clustering. Par conséquent, les méthodes basées sur une grille peuvent être intégrées à d'autres méthodes de clustering telles que les méthodes basées sur la densité et les méthodes hiérarchiques [22].

2.5 Clustering et anonymat

Les études actuelles sur le " privacy preserving clustering" peuvent être classées en deux types , à savoir les approches basées sur la perturbation et les approches basées sur la calcul multiparté sécurisé (SMC)3.6.

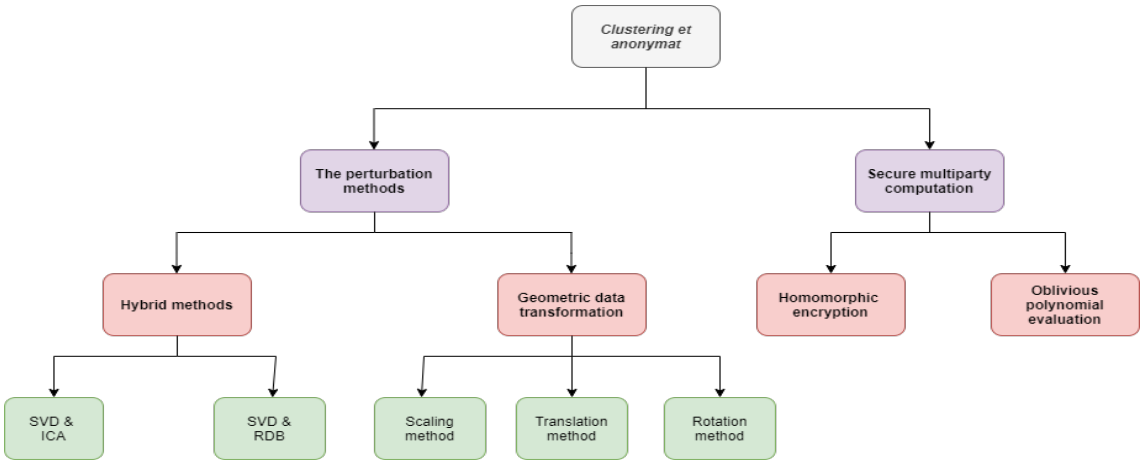


FIGURE 2.5 – Méthodes d"anonymat dans le cas du clustering

2.5.1 Les méthodes basées sur la perturbation des données

Ce sont des méthodes qui permettent de maintenir la confidentialité des données. Ces techniques modifient la valeur des attributs avec de nouvelles valeur sans autant changer la signification de base des données. On distingue [23].

Les méthodes de transformation de données géométriques

Ces méthodes transforment la valeur des attributs confidentielles par translation, scaling ou rotation.

Transformation des données par translation

Dans cette méthode, le terme de bruit appliqué à chaque attribut confidentiel est constant et peut être positif ou négatif. L'ensemble des opérations ne prend que la valeur Add correspondant à un bruit additif appliqué à chaque attribut confidentiel [23].

Pour illustrer le fonctionnement de la méthode TDP (The Translation Data Perturbation), considérons l'exemple de base de données relationnelle du tableau 2.1a. Dans cet exemple, on supprime les identifiants. Supposons que nous souhaitions regrouper des individus en fonction des attributs *Âge* et *Salaire*, mais les attributs sont confidentiels. Pour ce faire, nous appliquons la méthode TDP. Le vecteur de bruit uniforme pour cet exemple est N =(Add, -3, Add, 5000). Le tableau 3.1 montre la base de données déformée, et les points avant et aprés la distorsion comme c'est montré sur la figure 2.6.

Occupation	**City**	**Age**	**Salary**
Student	Edmonton	29	48,000
Executive	Calgary	38	72,000
Professor	Edmonton	34	51,000
Lawyer	Vancouver	43	65,000
Dentist	Victoria	42	60,000
Nurse	Toronto	48	53,000

(a) table de données simple [23]

Occupation	**Age**	**Salary**
Student	26	53,000
Executive	35	77,000
Professor	31	56,000
Lawyer	40	70,000
Dentist	39	65,000
Nurse	45	58,000

(b) A translation data perturbation [23].

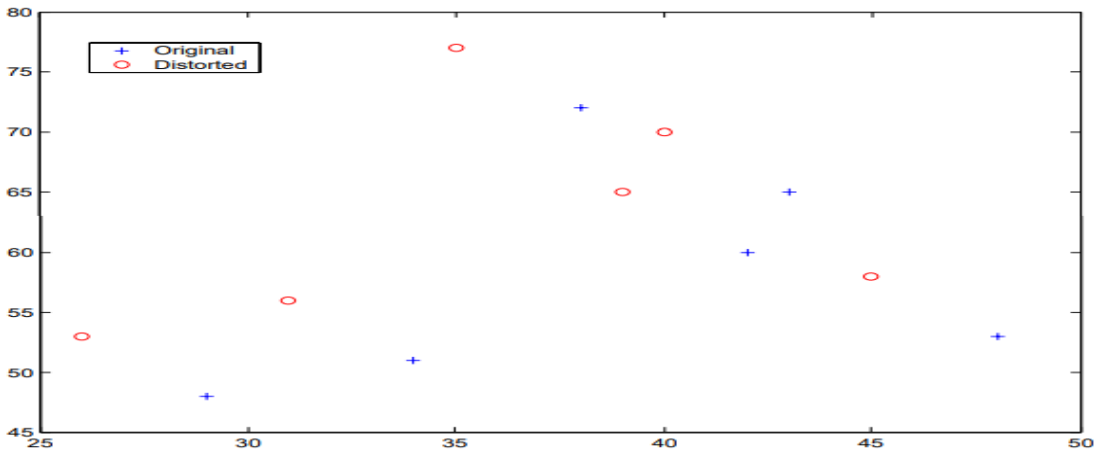


FIGURE 2.6 – Représentation des points avant "+" et aprés "o" la perturbation [23].

Avantages

— Fonctionne indépendamment pour chaque valeur capturée (adapté à la collecte de données).

— préserve les propriétés statistiques après reconstruction de la distribution d'origine.

Inconvénients

— limite l'utilité des données à l'utilisation de distributions agrégées.
— le masquage des valeurs extrêmes (telles que les valeurs aberrantes) nécessite de grandes quantités de bruit, ce qui dégrade considérablement les données.
— des techniques de réduction du bruit peuvent être utilisées pour estimer avec précision les valeurs individuelles d'origine.

Transformation des données par scaling

Dans cette méthode, le terme de bruit appliqué à chaque attribut confidentiel est constant et peut être positif ou négatif. L'ensemble des opérations ne prend que la valeur Multi correspondent à un bruit multiplicatif appliqué à chaque attribut confidentiel [23].

Pour illustrer le fonctionnement de la méthode SDP (The Scaling Data Perturbation), considérons l'exemple de base de données relationnelle du tableau 3.2. Dans cet exemple, nous souhaitons regrouper les individus en fonction des attributs Âge et salaire. Le vecteur de bruit uniforme pour cet exemple est N = (multi, 0,94, Mult, 1,035). Le tableau 3.3 montre la base de données déformée, et les points avant et aprés la distorsion sont montrés sur la figure 2.7.

Occupation	**City**	**Age**	**Salary**
Student	Edmonton	29	48,000
Executive	Calgary	38	72,000
Professor	Edmonton	34	51,000
Lawyer	Vancouver	43	65,000
Dentist	Victoria	42	60,000
Nurse	Toronto	48	53,000

(a) table de donnée simple [23].

Occupation	**Age**	**Salary**
Student	27	49,680
Executive	35	74,520
Professor	32	52,785
Lawyer	40	67,275
Dentist	39	62,100
Nurse	45	54,855

(b) A scaling data perturbation [23].

Avantages

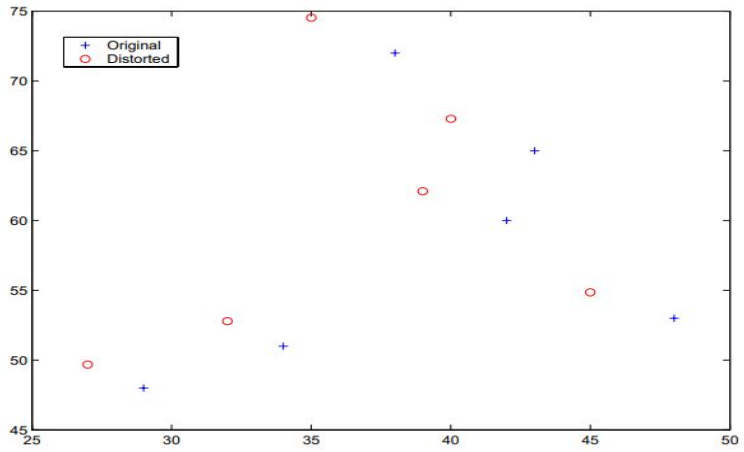


FIGURE 2.7 – Représentation des points avant "+" et aprés "o" la perturbation [23].

— Plus efficace que le bruit additif pour préserver la vie privée. étant donné que la reconstruction des valeurs individuelles d'origine est plus difficile
— Fonctionne indépendamment pour chaque valeur capturée (adapté à la collecte de données)
— Préserve les propriétés statistiques après reconstruction de la distribution d'origine

Inconvénients

— Limite l'utilité des données à l'utilisation de distributions agrégées
— Le masquage de valeurs extrêmes (telles que des valeurs aberrantes) nécessite de grandes quantités de bruit, ce qui dégrade gravement les données.

Transformation des données par Rotation

Cette méthode fonctionne différemment des méthodes précédentes. Dans ce cas, le terme de bruit est un angle. L'angle de rotation, mesuré dans le sens horaire, est la transformation appliquée aux observations des attributs confidentiels. L'ensemble des opérations prend uniquement la valeur Rotate qui identifie un angle de rotation commun entre les attributs Ai et Aj. Contrairement aux méthodes précédentes, RDP(The Rotation Data Perturbation) peut être appliqué plus d'une fois à certains attributs confidentiels [23].

Par souci de simplicité, nous illustrons le fonctionnement de la méthode RDP dans un espace discret 2D. Considérons l'exemple de base de données relationnelle du tableau 2.3a. Dans cet exemple, nous souhaitons regrouper les individus en fonction des attributs Âge et salaire. Le vecteur de bruit uniforme pour cet exemple est N = (Age ® Sal, 13,7). Le

tableau 2.3b montre la base de données déformée, et les points avant et aprés la distorsion sont montrés sur la figure 2.8.

Occupation	**City**	**Age**	**Salary**
Student	Edmonton	29	48,000
Executive	Calgary	38	72,000
Professor	Edmonton	34	51,000
Lawyer	Vancouver	43	65,000
Dentist	Victoria	42	60,000
Nurse	Toronto	48	53,000

(a) table de donnée simple [23].

Occupation	**Age**	**Salary**
Student	40	39,766
Executive	54	60,951
Professor	45	41,496
Lawyer	57	52,966
Dentist	55	48,354
Nurse	59	40,123

(b) A rotation data perturbation [23].

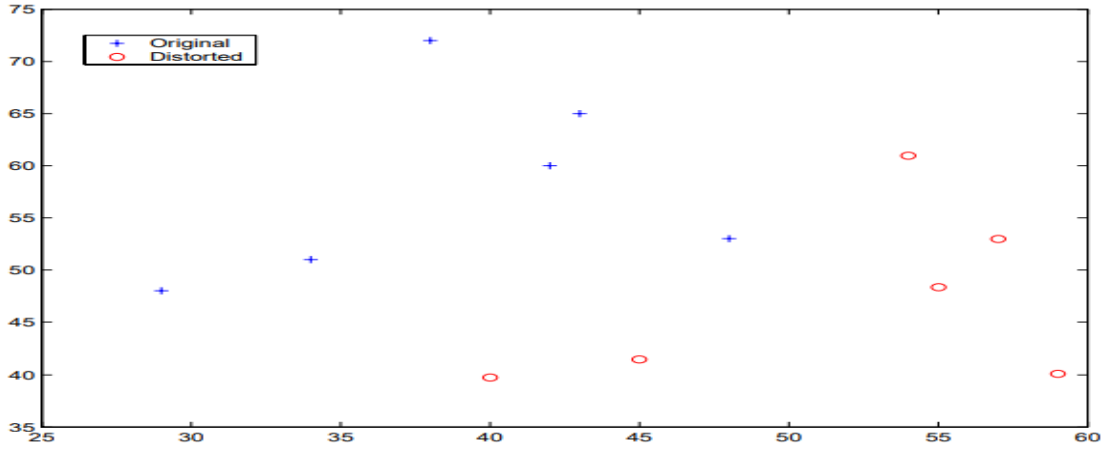


FIGURE 2.8 – Représentation des points avant "+" et aprés "o" la perturbation [23].

Avantages

— la méthode est indépendante de tout algorithme de clustering..
— la méthode a une base mathématique solide.
— Il ne repose pas sur des hypothèses d'inadaptabilité issues de l'algèbre et ne nécessite pas d'opération intensive du processeur.

Inconvénients

— Dans un espace discret 3D ou supérieur, deux variables sont affectées et les autres restent inchangées.
— appliquer une ou plusieurs transformations de rotation pour s'assurer que tous les attributs confidentiels sont déformés pour préserver la vie privée.

Les méthodes hybrides

Pour améliorer les performances de la perturbation de données unique basée sur SVD, deux méthodes de transformation de données hybrides sont proposées pour le clustering. Dans la première méthode hybride, la SVD et la perturbation des données de rotation sont utilisées en combinaison pour obtenir l'ensemble de données déformé. Dans la deuxième méthode hybride, la SVD et l'analyse en composantes indépendantes sont utilisées en combinaison pour obtenir l'ensemble de données déformé.

Décomposition en valeurs singulières (SVD)

Singular value Dicomposition (SVD) est l'une des méthodes familières [24]. La dimensionnalité de l'ensemble de données d'origine peut être réduit par SVD et également utilisée comme méthode de distorsion des données. L'ensemble de données d'origine A est représenté sous la forme d'une matrice n × m. Les objets de données sont représentés sous forme de lignes et les attributs sont représentés sous forme de colonnes. La décomposition en valeurs singulières est une méthode plus générale qui factorise tout nombre n × m matrice A de rang r en un produit de trois matrices, de telle sorte que :

$$A = UWV^T. \tag{2.1}$$

De la formule ci-dessus, U est une matrice orthonormée n × n, W est une matrice diagonale n × m dont les entrées diagonales non négatives (les valeurs singulières) sont dans l'ordre décroissant [24], où le nombre d'entrées diagonales non nulles est de rang(A) [25], et V^T est une matrice orthonormée m × m. En raison de la disposition des valeurs singulières dans la matrice, la transformation SVD a la propriété que la variation maximale des objets est

prise dans la première dimension et que la plupart des variations restantes sont capturées dans la deuxième dimension, et ainsi de suite [24]. L'approximation de rang k de A_Kà la matrice A peut être définie comme :

$$A_K = U_K W_K V_K^T. \tag{2.2}$$

A partir de la formule ci-dessus, on a un paramètre k tel que $0 \leq$ k $\leq$ Rang(A) [25], U_K contient les k premières colonnes de U, $W_K$contient les premières valeurs singulières non nulles, et V_K^T contient les k premières lignes de V^T. Avec k étant généralement petit, des valeurs plus élevées de k se traduisent par une meilleure utilité des données mais une protection de la vie privée pire. Ainsi, une matrice transformée avec une dimension beaucoup plus faible peut être construite pour représenter fidèlement la matrice d'origine [24].

Avantages

— protège efficacement les données privées des individus et conservent les informations importantes pour le clustering.

Inconvénients

— Ce n'est pas non plus tout à fait exact.

— SVD peut etre couteux en calcul.

L'analyse en composantes indépendantes (ICA)

L'ICA (Independent Component Analysis) est une méthode statistique permettant de transformer un ensemble de données complexe en sous-parties indépendantes. Le modèle ICA représente l'ensemble de données observé X comme une combinaison linéaire de matrice de mélange et de matrice aléatoire.

$$X = AS. \tag{2.3}$$

Dans la formule ci-dessus, S est une matrice aléatoire m $\times$ n représentant des coefficients ICA dont les valeurs sont supposées indépendantes et A est une matrice de mélange n $\times$ n. Avant d'appliquer l'algorithme ICA, les données doivent être centrées et blanchies. Le centrage est le processus consistant à soustraire sa moyenne des variables et à la convertir en variables moyennes nulles. Le merlan est une stratégie de prétraitement qui transforme

les composants de la matrice de données en non corrélés et la variance égale à l'unité. Les coefficients ICA de la matrice sont la représentation de codage clairsemé des données originales Les coefficients ICA peuvent être considérés comme un codage clairsemé des données originales. Dans le codage clairsemé, les éléments avec des valeurs abstraites plus grandes contiennent des informations plus importantes. Les éléments avec des valeurs abstraites plus faibles sont moins importants et sont donc souvent identifiés comme du bruit. Les composantes indépendantes sont des variables latentes, ce qui signifie qu'elles ne peuvent pas être observées directement et que la matrice de mélange est supposée inconnue. Seule X, la matrice observée est connue et estimer à la fois A et S en l'utilisant. Puis, après avoir estimé la matrice de mélange A, calculer son inverse, disons W pour obtenir les composantes indépendantes en utilisant la formule suivante.

$$S = WX. \tag{2.4}$$

Dans notre méthode, nous adoptons cette hypothèse. Nous considérons que les éléments de B avec de grandes valeurs abstraites représentent les tendances générales des données et contiennent des informations importantes pour l'exploration de données et que les éléments de B avec de petites valeurs abstraites ne sont pas importants pour l'exploration de données [24].

Les avantages

— facile à mettre en œuvre.

— elle fournit à la fois une visualisation optimale des variables et des données.

Les inconvénients

— Les représentations de ICA ne sont fiables que si la somme des pourcentages de variabilité associés aux axes de l'espace de représentation, est suffisamment élevée.

La méthode hybride-1 (SVD et RDP)

est proposée en tirant parti de deux techniques existantes SVD et perturbation de données de rotation pour optimiser la confidentialité fournie par la méthode de perturbation de données unique. L'ensemble de données d'entrée donné est prétraité en supprimant les attributs inutiles pour l'exploration de données et normalisé à l'aide de la normalisation du score z. L'ensemble de données est décomposé à l'aide de la perturbation de données SVD en trois matrices U ,W,

V^T. La matriceV^T ' est l'entrée des données de rotation perturbées et déformées en V^T '. L'ensemble de données final déformé est calculé comme un produit des matrices U, W, V^T ' [25].

Les avantages

— Elle atteint des précisions très satisfaisante par rapport aux autres méthodes.

Les inconvénients

— Limite de l'analyse du contenu

— méthode complexe

La méthode hybride-2(SVD & ICA) La sécurité fournie par le SVD de perturbation de données unique peut être renforcée par la méthode hybride proposée (SVD & ICA). Cette méthode est développée comme l'hybridation de la SVD et de l'analyse en composantes indépendantes. L'ensemble de données d'entrée donné est prétraité en supprimant les attributs inutiles pour l'exploration de données. L'ensemble de données est décomposé à l'aide de la perturbation de données SVD en trois matrices U ,W, V^T . La matrice V^T est l'entrée pour ICA, le modèle ICA exprime la matrice d'entrée VT comme une combinaison linéaire de matrices A et S et La matrice S constituée de composants indépendants est considérée comme une matrice déforméeV^T '. Le jeu de données final déformé est calculé comme un produit des matrices U, W,V^T ' [25].

Les avantages

— Peut gérer des données volumineuses avec une précision satisfaisante contrairement aux méthodes précédentes.

Les inconvénients

— méthode complexe.

2.5.2 Le calcul multipartit sécurisé

SMC (Secure Multi-party Computation), également appelé évaluation de fonction sécurisée (SFE), est un type de calcul préservant la confidentialité dans lequel deux parties ou plus calculent collectivement une fonction et reçoivent sa sortie sans qu'aucune partie n'apprenne les entrées privées des autres parties

Homomorphic Encryption

Le cryptage homomorphe permet de calculer avec des données cryptées et d'obtenir les mêmes résultats avec la version simple des données. La caractéristique la plus importante de ce type de schéma cryptographique est de préserver la confidentialité des données sensibles car elles permettent de travailler sur les données cryptées au lieu de leur forme simple. Les schémas de cryptage homomorphes peuvent également être utilisés pour connecter différents types de services sans risquer l'exposition de données sensibles [26]. La propriété homomorphe des schémas de cryptage homomorphes permet d'effectuer certaines opérations sur les données cryptées et de fournir des résultats cryptés. Après le décryptage, les mêmes résultats peuvent être obtenus avec des opérations effectuées en texte brut. Pour deux messages m1 et m2, dans l'équation, montre un schéma HE qui prend en charge toute opération $c1 = Enc(pk, m1), c2 = Enc(pk, m2)$ $m1m2 = Dec(sk, c1c2)$ Dans lequel Enc(·) et Dec(·) sont les algorithmes de cryptage et de décryptage ; c1 et c2 sont les cryptages de m1 et m2 [27]. Un schéma appelé additivement (ou multiplicativement) homomorphe si le résultat de calcul de deux variables cryptées est le même que celui des variables d'origine. $[x][y] = [x + y]and[x][y] = [x\mathring{u}y]$ pour l'addition et la multiplication, respectivement, où [m] désigne le cryptage de certains textes en clair m. Les symboles et désignent respectivement les opérations d'addition et de multiplication homomorphes dans l'espace du texte crypté. En d'autres termes, si un schéma de cryptage est additivement homomorphe, alors le cryptage suivi d'une addition homomorphe est égal à l'addition suivie d'un cryptage [28].

Les avantages

- les organisations peuvent établir un niveau de sécurité des données plus élevé sans perturber les processus métier ou les fonctionnalités des applications. Ces organisations peuvent garantir la confidentialité des données, tout en tirant des renseignements de leurs données sensibles.
- Les cas d'utilisation du cryptage homomorphe incluent la protection de la charge de travail dans le cloud (ou "lift and shift" vers le cloud), l'analyse agrégée (cryptage préservant la confidentialité), la consolidation de la chaîne d'approvisionnement des informations (contenant vos données pour atténuer le risque de violation), et l'automatisation et l'orchestration (exploitation et déclenchement des données cryptées pour la communication de machine à machine).

Les inconvénients

— Pour certains algorithmes cryptographiques, après avoir appliqué l'algorithme de cryptage sur des données en texte brut, la taille du texte chiffré est plus comparable à celle du texte brut d'origine. La raison peut être due à une procédure de rembourrage. Ainsi, pour effectuer des calculs sur ces données cryptées, il faudra plus de temps de calcul.

— Le texte chiffré peut comporter des éléments de bruit qui deviennent relativement massifs avec les calculs de multiplication homomorphes ultérieurs, et seuls les textes chiffrés, dont l'estimation du bruit reste dans une certaine valeur de seuil, peuvent être déchiffrés avec précision .

Oblivious ploynomial evaluation

Est un protocole impliquant deux parties, un expéditeur dont l'entrée est un polynôme P, et un récepteur dont l'entrée est une valeur α. A la fin du protocole le récepteur apprend $P(\alpha)$ et l'expéditeur n'apprend rien. Il est utilisé comme mécanisme cryptographique pour assurer la sécurité des données privées [29].

Les avantages

— Ces protocoles assurent la confidentialité contre les expéditeurs semi-honnêtes ou malveillants.

— Les protocoles garantissent qu'un récepteur malveillant apprend au plus une seule équation linéaire des coefficients du polynôme.

Les inconvénients

— Ces protocoles assurent la confidentialité contre les expéditeurs semi-honnêtes ou malveillants.

— Les protocoles garantissent qu'un récepteur malveillant apprend au plus une seule équation linéaire des coefficients du polynôme.

2.6 Etude comparative

Dans cette section nous allons présenter un résumé sur les différentes méthodes d'anonymisation des données appliquées pour le cas du clustering :

Les méthodes	**Type**	**Stratégies**	**Point fort**	**Point faible**
géométriques	Translation	-les données sont des perturbations en ajoutant du bruit.	-les distances entre les points de données sont préservées.	- Non disponible pour tous les types de données.
	Scaling	-les données sont des perturbations en multipliant le bruit	-Préserver les propriétés statistiques après la reconstruction des distributions.	-une faible confidentialité. -Les valeur rélles des données sous clustering peut être estimée.
	Rotation	les données sont des perturbations en roting noise	C'est efficace et précis Pour tous les algorithmes de clustering.	- Les données bruitées et le clusters déséquilibré rendent cette méthode difficiles à répondre aux besoins pratiques
hybrid	Singular value decomposition (SVD)	décomposition en valeurs singulières	-il est utilisé pour trouver les informations sans importance pour l'extraction de données et les supprimer pour protéger la vie privée	-Il est lent en termes de temps
	Independent Component Analysis (ICA)	Composante indépendante de données.	- il existe des possibilités d'améliorer encore la sécurité sans perdre la mesure de la précision	le temps de calcul peut augmenter avec le nombre d'échantillons

	oblivious polynomial evaluation		-Applicabilité à tout type d'attribut.	-Ils sont très complexes en calcul.
Secure Multi-party Computation	homomorphic encryption	Encryption	- Fournir une meilleure protection de la vie privée. - Peut être appliqué à divers types de données.	-calculs importants en raison du processus de cryptage et décryptage, en particulier pour les données de grande taille -Nécessite un travail supplémentaire difficile pour préparer un générateur de nombres aléatoires qui résiste à attaque par flux d'entropie

TABLE 2.4 – les méthodes de clustering

2.7 Conclusion

La technologie de clustering est l'une des techniques d'exploration de données les plus couramment utilisées. Appliqué à divers domaines où les données arrivent sous la forme d'un flux. Dans ce chapitre nous avons étudié différentes méthodes liées à cette technique dans ce chapitre. Nous avons présenté 4 méthodes de clustering, les méthodes hiérarchiques et de partitionnement et les méthodes basé sur la densité et sur la grille. Ainsi, Nous avons détaillé leurs principes de fonctionnement et présenté leurs Points faibles et leurs points forts.

Chapitre 3

Proposition et validation

3.1 Introduction

Comme on l'a vu dans le chapitre précédent, il existe plusieurs méthodes pour le problème de privacy preserving clustering. Dans ce chapitre nous allons présenter notre méthode que nous avons proposés et qui permet de perturber un ensemble de données pour le rendre anonyme tout en gardant l'utilité des données.

Nous allons présenter les détails de notre proposition, son architecture, notre méthode pour l'ajout du bruit et les paramètres sur lesquels on s'est basé pour évaluer notre proposition. Nous présenterons aussi les différentes bibiothéques et le langage de programmation utilisé pour notre implémentation et validation et nous présenterons quelques résultats que nous avons obtenus.

3.2 Architecture Proposé

La figure suivante représente l'architecture de notre approche : Dans notre architecture, on effectue en premier lieu un prétraitement des données pour nettoyer notre dataset,ensuite on remplace les données originales de notre dataset par des nouvele valeurs synthétiques générées par la méthode CTGAN , pour avoir un nouveau dataset perturbé avec de nouveaux enregistrement qui sera ensuite partagé pour pouvoir effectuer sur ce dernier un clustering.

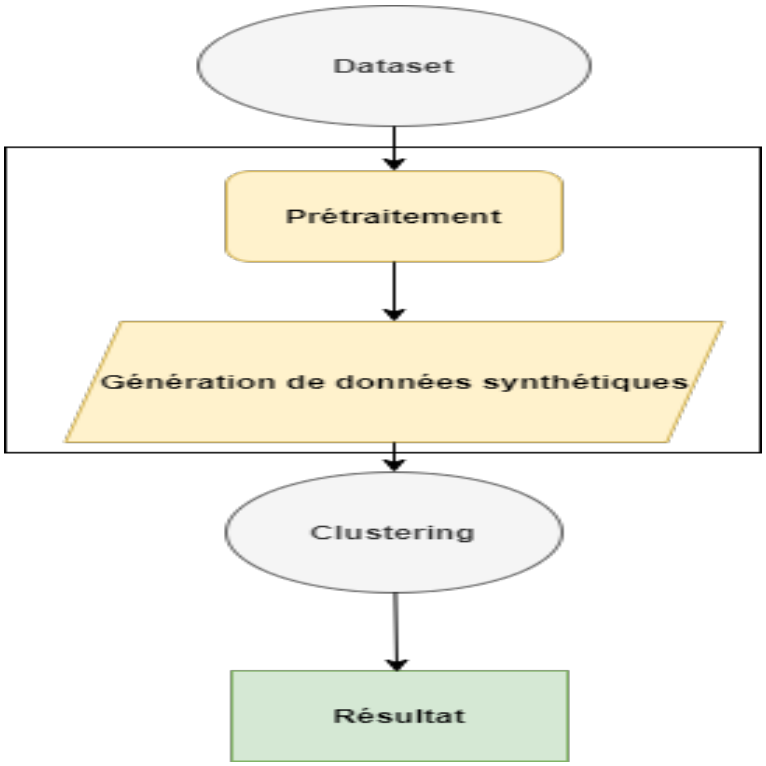


FIGURE 3.1 – Architecture de programme

3.3 Visualisation

La visualisation des données est la représentation graphique des informations et des données. À l'aide d'éléments de visualisation tels que des graphiques, des diagrammes et des cartes, pour faciliter la compréhension et l'obtention d'informations à partir des données par le cerveau humain. L'objectif principal de la visualisation des données est d'afficher et de comprendre les tendances, les valeurs aberrantes et les modèles de données.

Maintenant, nous prenons et regardons quelques exemples sur nos données :

— Afficher le nombre d'observations dans chaque casier catégoriel à l'aide de barres.

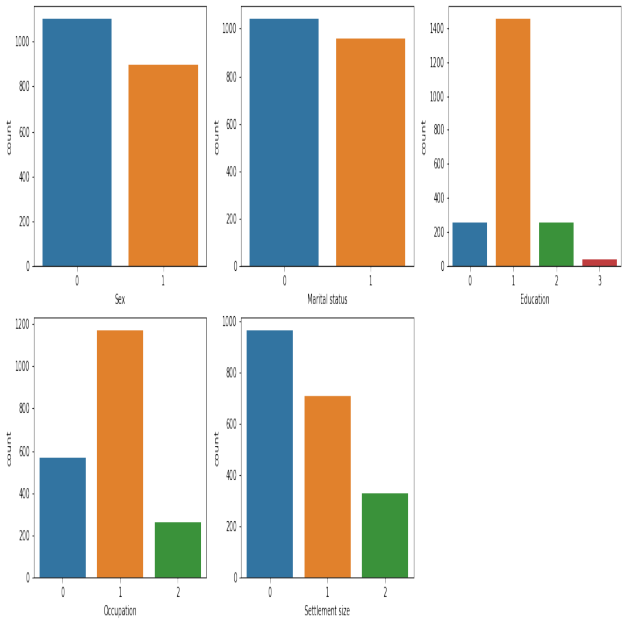


FIGURE 3.2 – Visualisez le nombre total de colonnes de chaque type dans les données.

Explication : D'après le résultat obtenu en figure 3.2 ci-dessus, 1086 clients sont des hommes et 914 sont des femmes, 1007 clients sont célibataires et 993 sont non célibataires, 287 clients dont le niveau d'éducation est inconnu, 1386 clients dont le niveau d'éducation est secondaire, 287 sont à l'université et 36 clients sont en études supérieures.

— **Un histogramme** est un outil de visualisation classique qui représente la distribution d'une ou plusieurs variables en comptant le nombre d'observations qui se trouvent dans des bacs discrets [40].

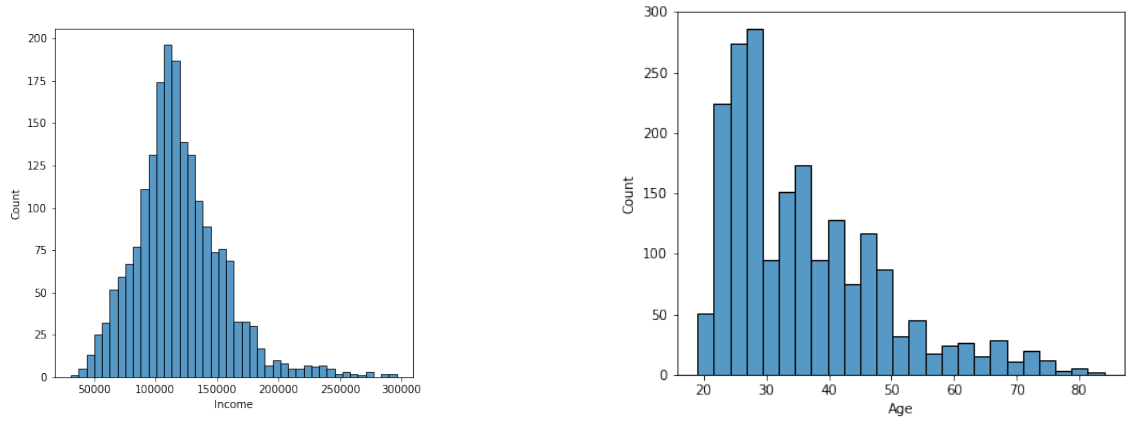


FIGURE 3.3 – Histogramme de l'Age et Income.

Explication : D'après le résultat obtenu dans l'histogramme ci-dessus, la distribution des données de l'age et de Income est gaussienne.

3.4 Pré-traitement

Le Pré-traitement des données consiste à décrir tout type de traitement primaire effectué sur des données brutes pour les préparer à d'autres opérations de traitement avancées. Les technologies de prétraitement des données visent à convertir des données réelles dans un format compréhensible qui facilite et augmente l'efficacité des opérations de traitement requises, nettoyer les données des valeurs incorrectes, supprimer les lignes et les colonnes vides, ...

3.4.1 Affichage des données

la figuré suivante renvoie les cinq premières lignes du Dataset.

```
data.head()
```

	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
0	10000001	0	0	30	0	195925	2	2
1	10000002	1	1	29	1	41424	0	0
2	10000003	0	0	36	1	163240	2	2
3	10000004	1	1	24	1	86473	0	0
4	10000005	1	1	21	1	66404	0	0

FIGURE 3.4 – Affichage de données.

3.4.2 Affichage des informations

La figure suivante représente des informations sur le dataset.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ID               2000 non-null   int64
 1   Sex              2000 non-null   int64
 2   Marital status   2000 non-null   int64
 3   Age              2000 non-null   int64
 4   Education        2000 non-null   int64
 5   Income           2000 non-null   int64
 6   Occupation       2000 non-null   int64
 7   Settlement size  2000 non-null   int64
dtypes: int64(8)
memory usage: 125.1 KB
```

FIGURE 3.5 – Affichage des informations.

Dans le FIGURE 3.5 est affiché les colonnes de notre dataset, le type de chaque colonne, le nombre de valeurs nulls, ... D'aprés cette figure on constate qu'il n'ya aucun point de données manquant ou null.

3.4.3 Les valeurs manquantes

La figure suivante montre que les colonnes ne contiennent pas des valeurs manquantes.

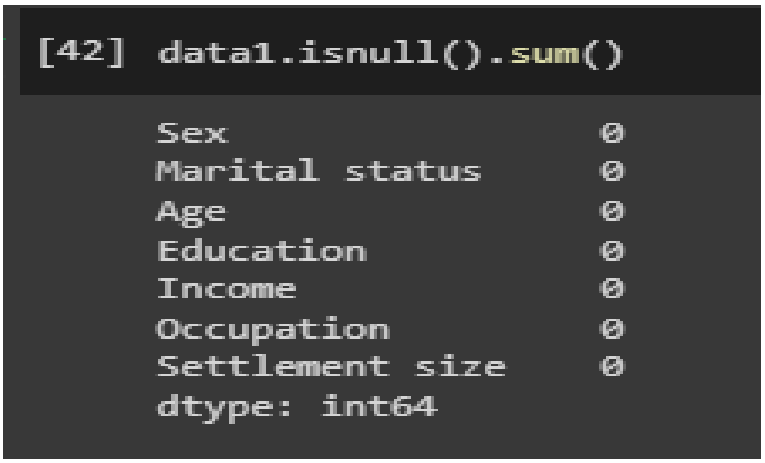


FIGURE 3.6 – Les valeurs manquantes de chaque colonne.

3.4.4 Nombre les lignes de duplication

La figure suivante montre que lignes ne contiennent aucune duplication.

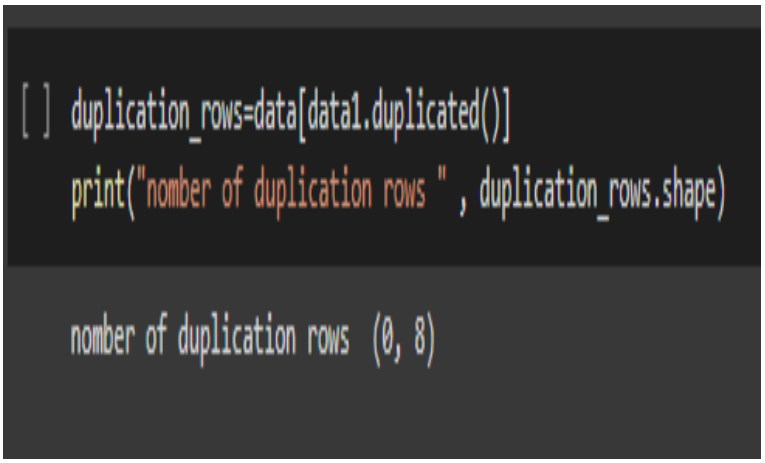


FIGURE 3.7 – Nombre les lignes de duplication .

3.4.5 PCA

La PCA est une tâche de pré-traitement non supervisée qui est effectuée avant d'appliquer un algorithme de ML. Dans la figure 3.8 nous choisissions le nombre de composants principaux de façon à pouvoir expliquer 91 %de la dispersion initiale des données (par l'intermédiaire du ratio explicatif de la variance). Ici, cela signifie de conserver 4 composants principaux.

figure suivante représente les resultat de l'ACP.

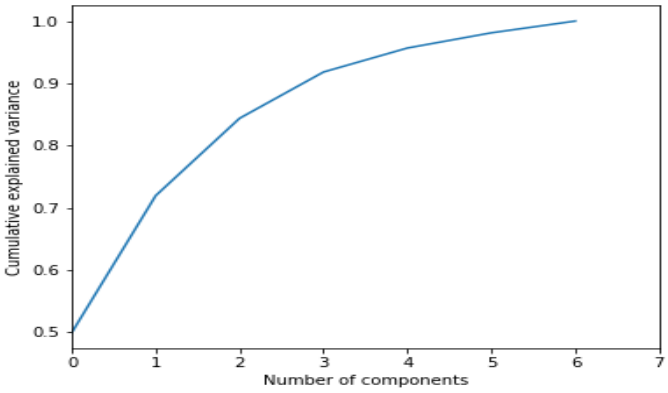


FIGURE 3.8 – Le nombre de composantes par rapport à la variance.

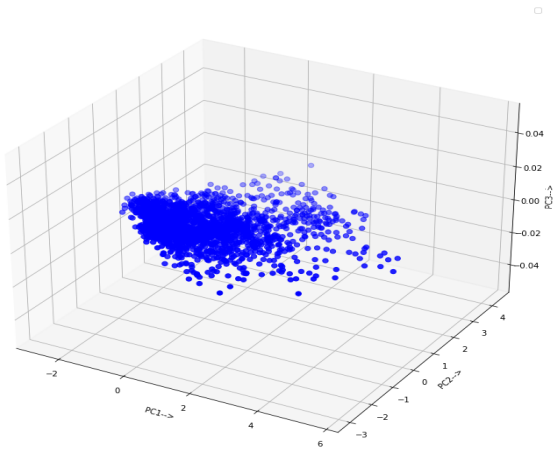


FIGURE 3.9 – Les résultats de l'APC 3D.

3.5 Proposition

3.5.1 GAN

Un réseau neuronal générateur (G) et un réseau neuronal discriminateur (D) sont les deux composants fondamentaux des réseaux antagonistes génératifs (GANs), comme montré par la figure 3.11 . Ces deux réseaux neuronaux sont toujours en compétition l'un avec l'autre pour améliorer la précision de leurs prédictions afin de générer de nouvelles instances synthétiques de données. Le générateur (G) apprend à créer de fausses données grâce au feedback du discriminateur. Son but est d'amener le discriminateur à classer sa

sortie (fake data) comme réelle.

Pour entraîner le générateur, il doit l'intégrer étroitement au discriminateur. L'entraînement consiste à prendre une entrée aléatoire (Random noise), puis la transformer en une instance de données, et l'envoyer au discriminateur et à recevoir une classification, ainsi qu'à calculer la perte du générateur, qui pénalise un jugement correct par le discriminateur.

Tandis que, le discriminateur (D) est un classificateur qui examine les données fournies par le générateur, et tente d'identifier s'il s'agit de fausses données générées ou de vraies données.L'apprentissage est effectué à l'aide d'instances de données réelles, utilisées comme exemples Positifs, tandis que les instances de données fictives générées par le générateur sont utilisées comme exemples négatifs.

Ce discriminateur utilise une fonction de perte (loss) qui pénalise une classification incorrecte d'une instance de données fausse comme réelle ou d'une instance réelle comme fausse. À chaque cycle de formation, le discriminateur met à jour les poids de son réseau neuronal, et s'améliore de plus en plus dans l'identification des instances de données fausses [30].

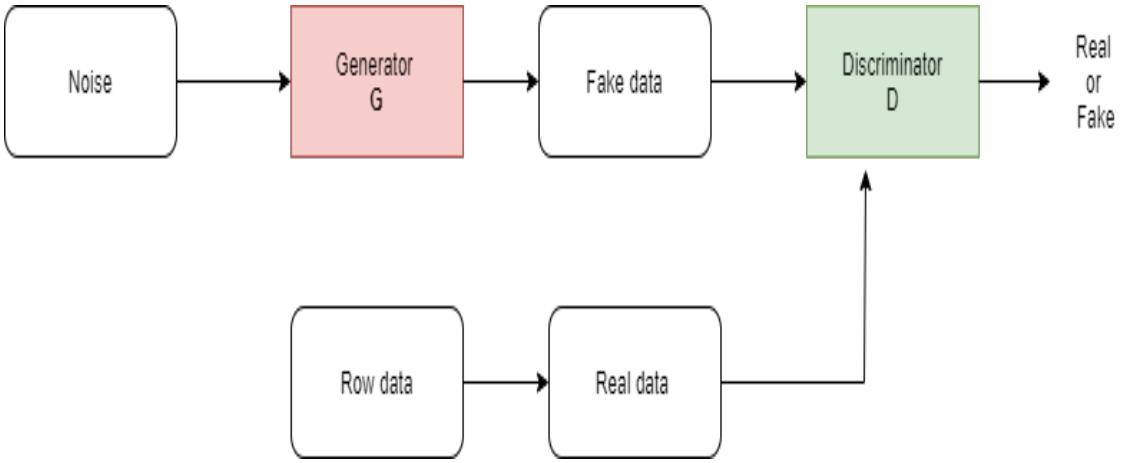


FIGURE 3.10 – Architecture de GAN .

3.5.2 Génération de données synthétiques

Même si GAN a prouvé sa capacité supérieure à générer des données, il ne peut pas générer complètement tous les types de données. Par exemple, GAN peut difficilement modéliser toutes les distributions dans des données tabulaires lorsque les variables indépendantes du tableau ont des distributions différentes. De plus, si les données incluent des informations catégorielles représentées par un vecteur un-chaud, GAN ne peut pas garantir les propriétés du vecteur un-chaud.

CTGAN a été proposé pour résoudre ce problème à l'aide de deux nouvelles techniques. La première technique est la normalisation spécifique au mode N (0, 1) pour une variable continue constituée de valeurs flottantes. CTGAN utilise un modèle de mélange gaussien variationnel comme modèle de distribution unitaire pour estimer le nombre de distributions pour chaque variable et normaliser les valeurs des variables en fonction des distributions estimées. Ensuite, CTGAN utilise ces valeurs codées à la place des valeurs d'origine pendant l'entraînement. Lors de la génération de données artificielles après la formation, CTGAN transforme les données générées en échelle d'origine. La deuxième technique est une approche d'entraînement conditionnel pour gérer les fréquences déséquilibrées au niveau des catégories dans les variables catégorielles. Le déséquilibre de fréquence fait que le générateur GAN ne produit que quelques catégories qui apparaissent fréquemment dans les données. Par conséquent, ce problème devrait être résolu pour générer diverses données. L'approche d'apprentissage conditionnel aborde ce problème de la manière suivante : chaque colonne des données tabulaires et les variables catégorielles sont codées en vecteurs de condition. Ces vecteurs sont échantillonnés en fonction de la fréquence logarithmique des catégories pour s'assurer que les niveaux catégoriques rares sont échantillonnés uniformément [30].

Dans notre proposition, nous appliquons cette méthode pour remplacer les valeurs originales de notre dataset par de nouvelles valeurs synthéthiques générées par la méthode CTGAN tout en essayons de garder l'utilité des données originales de notre dataset et ainsi rendre la ré-identification des individus difficile.

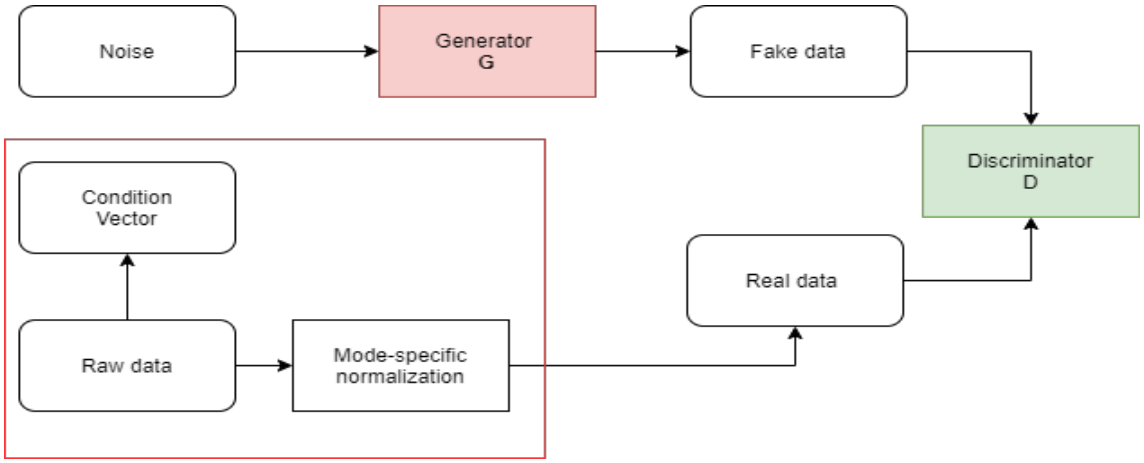


FIGURE 3.11 – Architecture de tabulaire conditionnel GAN .

3.6 Evaluation

3.6.1 Dataset utilisé

Le Dataset que nous avons choisi se compose d'informations sur le comportement d'achat de 2000 individus d'une zone donnée lors de la saisie d'un store physique "FMCG". Toutes les données ont été collectées via les cartes de fidélité qu'utilisent les individus à la caisse. L'identité des individus a été supprimé.

Attribut	**Explication**
ID	Identifiant unique pour chaque client
Sexe	Sexe biologique d'un client
Marital status	Etat civil d'un client
Age	L'âge du client en années. calculé comme l'année en cours moins l'année de naissance du client au moment de la création de l'ensemble de données
Eduction	Niveau d'éducation du client
Income	Revenu annuel auto déclaré en dollars américains du client
Occupation	Catégorie d'occupation du client
Settlement size	La taille de la ville dans laquelle vit le client

TABLE 3.1 – Explication des données de dataset.

3.6.2 Environnements et outils de développement

Langage utilisé

Python

Python est un langage portable, dynamique, extensible, gratuit, qui permet sans l'imposer une approche modulaire et orientée objet de la programmation. C'est est un langage de programmation généraliste, facile à apprendre et rapide à mettre en œuvre .Il est perfermonte et open source[31].

Plateforme et environnement de développement

Colab

Google Colab ou Colaboratory est un service cloud, offert par Google (gratuit), basé sur Jupyter Notebook et destiné à la formation et à la recherche dans l'apprentissage automatique. Cette plateforme permet d'entraîner des modèles de Machine Learning directement dans le cloud. Sans donc avoir besoin d'installer quoi que ce soit sur notre ordinateur à l'exception d'un navigateur [32].

Kaggle

Kaggle est une platforme web qui accuielle la plus grande communauté de data science au monde. Cette plateforme propose des projets intéressnts ou les contributeurs peuvent apprende et s'exercer [33].

Bibliothèques Utilisés

Pandas

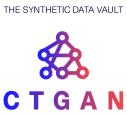
Pandas est une bibliothéque python open source permettant de l'analyse,exploration et manipulation de données [34].

Numpy

Numpy est une bibliothéque de python utilisé dans les domaines de la science

et l'ingénierie.Elle contient des tableaux multidimensionnels et des structures de données matricielles [35].

Ctgan



Ctgan est une collection de générateurs de données synthétiques basés sur le Deep Learning pour les données de table unique, qui sont capables d'apprendre des données réelles et de générer des clones synthétiques avec une haute fidélité [36].

Matplotib

Matplotib est une bibliothéque complète pour créer des visualisations statiques,animées et interactives en python [37].

Sklearn

Sklearn est une bibliothéque de python open source destinée a l'apprentissage automatique .Cette bibliothéque qui est une grande partie écrite en python ,s'appuie sur Numpy ,Scipy et Matplotlib [38].

Table Evaluator

Table Evaluator est une bibliothèque permettant d'évaluer la similitude d'un ensemble de données synthétisé avec des données réelles. En d'autres termes, il essaie de donner une indication de la réalité de vos fausses données [39].

Méthode de clustering utilisée

De nombreux algorithmes ont été proposés dans le domaine clustering. Leurs principes déffèrent selon plusieurs critères :le type de données à traiter, la complexité de l'algorithme , les mesures de distances utilisée ect [41].

K-means

L'algorithme K-means est un algorithme populaire de regroupement de données. Il permet de regrouper en K clusters distincts les observations du data set. Ainsi les données similaires se retrouveront dans un même cluster [42]. algorithme consiste à sélectionner

aléatoirement k objets qui représente les centroïdes initiaux. Un objet est assigné au cluster pour lequel la distance entre l'objet et le centroïde est minimale. La formule suivante représente la variance des clusters :

$$V = \sum_{j} \sum_{x_i \to c_j} d(c_j, x_i)^2 \tag{3.1}$$

Avec :

c_j : Le centre du cluster (le centroïd).

x_i : la ième observation dans le cluster ayant pour centroïd.

d(c_j,x_i) : La distance (euclidienne ou autre) entre le centre du cluster et le point.

Les avantages

- très facile pour l'implémentation.
- temps de calcul acceptable.
- K-means convient à un grand nombre d'ensembles de données.
- la complixité de k means est O(n).

Les inconvénients

- Spécifier le nombre de clusters K, avant l'application de l'algorithme.
- Il n'est pas adapté aux données non numériques.
- Le résultat dépend du tirage initial des centroïds

3.6.3 Résultats expérimentaux

Entropie

Définition

L'entropie est une mesure de quantité d'information moyenne d'un ensemble d'évènements. L'entropie inférieure signifie un meilleur regroupement. L'entropie s'amplifie lorsque la vérité au sol des objets dans le cluster se diversifie davantage. L'entropie plus grande signifie que le clustering n'est pas bon. La quantité de désordre est trouvée en utilisant l'entropie. , l'entropie est calculé par la formule suivante [43] :

$$H(x) = -\sum_{i=1}^{n} P(i) \log(P(i)) \tag{3.2}$$

Tel que :

H(x) : entropie d'un seul cluster .

P(x_i) : est la probabilité d'associée à l'apparition de l'évènement i.

Résultat d'évaluation de l'enrtopy

le tableau suivant nous montre les résultats d'évaluation de l'entropy :

Algorithme	**Data originale**	**Data perturbé**
K-means	0.93	0.96

TABLE 3.2 – Résultat de entropy .

D'après les resultats précedentes montrés dans le tableau3.2 , L'entropie inférieure signifie un meilleure groupemment. n ya pas de différence entre dataset originale et dataset perturbé donc la qualtié d'information est perdue légèrement .

Misclassification

Définition

Misclassification est la perte d'informations peut être mesurée par le pourcentage de points de données légitimes qui ne sont pas bien classés après le processus de désinfection. une erreur de classification erronée Misclassification Erreur est définie pour mesurer la perte d'informations [44].Elle est calculé par la formule suivante :

$$M_E = \frac{1}{N} \times \sum_{i=1}^{k} (|Cluster_i(D)| - |Cluster_i(D')|) \tag{3.3}$$

ou :

- N : represente le nombre de point originale.
- k : est le nombre de clusters sous analyse.
- $|Cluster_i(D)|$:représente le nombre de points de données légitimes du ième cluster dans la base de données D.
- $|Cluster_i(D')|$:représente le nombre de points de données légitimes du ième cluster dans la base de données D'.

Résultat d'évaluation de misclassification

le tableau suivant nous montre les résultats d'évaluation de misclassification :

Dans le tableau 3.3 nous avons les résultats des erreurs de classifications (misclassification).Nous avons comparé l'analyse de cluster des ensembles de données originaux et

Algorithme	**Misclassification**
K-means	0.087

TABLE 3.3 – Résultat de misclassification

déformées. Notre téchnique donne des bonnes résultats.Dans le pire des cas,seulement %8 des points sont mal classées.

Quantifying Privacy

Définition

Est une mesure qui convient parfaitement pour mesurer la variance de la déffirence entre les valeurs réelles et perturbées. Quantifier la Confidentialité inférieure signifie que les enregistrements de dataset perturbé ne sont pas différents de les enregistrements de dataset d'origine. Cette mesure est donner per Var(X-Y) ou X reprsente un seul attribut original et Y l'attribut perturbé.Cette mésure peut étre échelle invariante avec respect a la variance de X en expriment la valeur comme [45] :

$$S_1 = Var(X - Y)/Var(X). \tag{3.4}$$

Résultat d'évaluation de Quantifying Privacy

le tableau suivant nous montre les résultats d'évaluation de Quantifying Privacy :

column	**Quantifying Privacy**
Sexe	1.99
Marital status	1.99
Age	1.00
Eduction	1.89
Income	1.00
Occupation	1.95
Settlement size	1.82

TABLE 3.4 – Resultat de Quantifying Privacy.

D'aprés les résultats représentés dans le tableau 3.4, la variance de chaque attribut est élevé donc la confidentialité est bonnes.

Ala fin voici le resultat de notre approche :

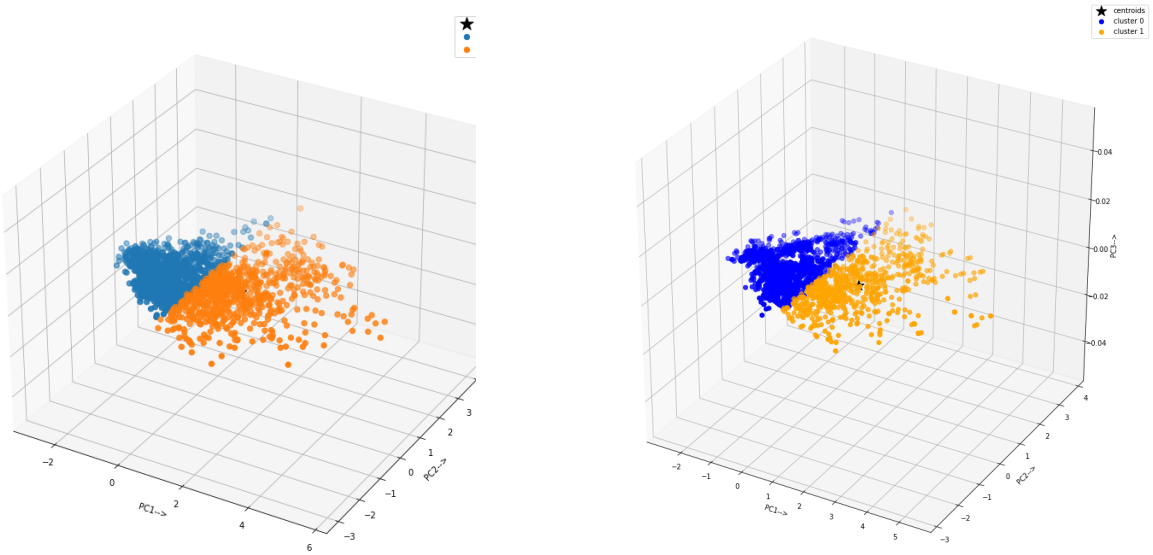


FIGURE 3.12 – Les résultats de deux datasets.

d'aprés la figure 3.12 en deduit qu'il n'ya pas une grande déffirence de classification des point dans les clusters.

3.7 Conclusion

Dans ce chapitre nous avons commencé par déffinir les différents environnements et outils de dévelopement utilisé pour l'implémentation de notre approche ainsi que nous avons représenté les mesures d'évaluation.En terminons avec une discussion des resultats obtenus de notre approche de Génération de données synthétiques en utilisons CTGAN .

Conclusion générale

L'algorithme de clustering diviser les données en groupes similaires. Lorsque les données proviennent de différentes sources où elles contiennent des données sensibles ,il est fortement recommandé de préserver la vie privée de ces données afin de résoudre le problème de la réidentification lors du clustering.

L'objectif de notre travail est de publier un dataset anonyme construit à partir d'un dataset original qui protége contre les risques de ré-identification, autrement dit, on réduit la probabilité de pouvoir trouver dans le dataset publié un individu présent dans la dateset original tout en gardant l'utilité des données du dataset original, et pour attiendre cet objectif nous avons proposé une méthode qui permet de perturber le dataset originale en utilsant la technique CTGAN en remplacement les données orginales par des données synthétiques générées par cette méthode pour répondre aux exigences de confidentialitéet préserver les caractéristiques générales de clustering.

Nos expériences ont démontré que notre méthode sont efficaces et fournissent des valeurs pratiquement acceptables pour équilibrer confidentialité et précision. La meilleure chose est qu'une cette méthode ne modifie pas la distance des points dans l'espace euclidien son résultat était 9%, ce qui réduit le problème critique de la mauvaise classification. Notre deuxième contribution fait référence à une métrique de performance qui quantifie la fraction de points de données enregistrés dans les groupes correspondants dans la base de données déformée.

travailler à un projet pareil, avec plusieurs technologies été bénéfique à plusieurs niveaux :

— Au niveaux technique, nous avons l'opportunité de maîtriser certains outils et technologies comme google colaboratory, kaggle. Ainsi, ca nous a permis de maîtriser le langage de programmation python.

— Au niveau personel, on a appris à gérer les données privées et on peut maintenir la confidentialité des données.

Perspectives

— Nous interessant d'utiliser notre travail pour l'anonymat des données destinées pour la classification.

— Nous voudrions tester cette approche avec d'autres algorithmes de clustering et avec des datasets différents.

— Nous suggérons la proposition de l'ajout des individus au dataset.

Bibliographie

[1] https://www.google.com/search?q=privacy+or+anonymity$&$client=firefox-b-d$&$ei=UMgIYvy8EMSX9u8PhfyQgA8$&$ved=0ahUKEwj8pKuUqPz1AhXEi, consulté le 11/02/2022

[2] https://www.fifosys.com/blog/security/ , consulté le 14/02/2022

[3] https://www.avast.com/fr-fr/c-identity-theft , consulté le 15/02/2022

[4] https://pcs.marsh.com/ca/fr/insights/thought-leadership/identity-theft.html, consulté le 17/02/2022

[5] Jean-Philippe Walter, Le profilage des individus à l'heure du cyberespace : un défi pour le respect du droit à la protection des données, téphanie Lacour (Ed.), La sécurité de l'individu numérisé. Réflexions prospectives et internationales, 2008

[6] L. Sweeney, . *K-anonymity : a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*,10(5) :557–570, 2002.

[7] A.Machanavajjhala, D.Kifer, J.Gehrke, M.Venkitasubramaniam,*l-Diversity : Privacy beyond k-anonymity* 2007.

[8] N.Li, T.Li, S.Venkatasubramanian,*t-Closeness : Privacy Beyond k-Anonymity and l-Diversity*,IEEE 23rd International Conference on Data Engineering,2007.

[9] Saravanan.k , Hemavathi.D, A Journey on *Privacy protection strategies in big data*, International Conference on Intelligent Computing and Control Systems ICICCS ,2017

[10] https://tel.archives-ouvertes.fr/tel-01783967/document , consulté le 17/02/2022

[11] P.Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6), 1010–1027, 2001 .

[12] https://cnpd.public.lu/content/dam/cnpd/fr/publications/groupe-art29\\wp216en.pdf , consulté le 23/02/2022

[13] J.Domingo-Ferrer, V.Torra, *Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. Data Mining and Knowledge Discovery*, 11(2), 195–212,2005

[14] G.Pillot, Anonymat et vie privée sur internet ,2018 .

[15] https://www.makeuseof.com/tag/3-undeniable-reasons-need-online-anonymity/, consulté le 24/02/2022.

[16] https://www.techno-science.net/glossaire-definition/Anonymat-sur-Internet.htmlref2, consulté le 26/02/2022.

[17] J. Han, M. Kamber, and J. Pei, Data Mining : Concepts and Techniques. San Mateo, CA, USA : Morgan Kaufmann, 2006.

[18] https ://analyticsinsights.io/le-clustering-definition-et-implementations/, consulté le 22/02/2022.

[19] https://tel.archives-ouvertes.fr/tel-00084828/document, consulté le 05/03/2022

[20] A.E. Ezugwu, A.M. Ikotun, O.O. Oyelade , L.Abualigah, J.O. Agushaka , C.I. Eke , A.A. Akinyelu .A comprehensive survey of clustering algorithms : State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects , Engineering Applications of Artificial Intelligence 110 , 2022 .

[21] Y.Zhaoa, S.K. Tarus, L.T. Yanga , J.Suna, Y.Gee , J.Wang . Privacy-preserving clustering for big data in cyber-physical-social systems : Survey and perspectives, information sciences 515, 2020 132-155.

[22] J.Han , M,Kamber , J.Pei . Data mining : concepts and techniques , 2012 .

[23] S.Oliveira, O.Zaïane. Privacy Preserving Clustering By Data Transformation. 2003

[24] M.N. Lakshmi , K.S. Rani. SVD based data transformation methods for privacy preserving clustering, International Journal of Computer Applications, Volume 78 – No.3, 39–43, September 2013.

[25] G.Li, Y.Wang. A privacy-preserving data mining method based on singular value decomposition and independent component analysis , Data Science Journal, Volume 9,16 , 124-132,February 2011.

[26] F.O.Catak , I.Aydin, O.Elezaj ,S.Y.Yayilgan .Practical Implementation of Privacy Preserving Clustering Methods Using a Partially Homomorphic Encryption Algorithm , 2020.

[27] W.Ren, X.Tong, J.Du, N.Wang, S.C.Li, G.Min, Z. Zhao, A.K.Bashir . Privacy-preserving using homomorphic encryption in Mobile IoT systems, Computer Communications,2020.

[28] A.Wood, K.Najarian, D.Kahrobaei. Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics. ACM Comput. Surv. 53, 4, Article 70 (July 2021), 35 pages,2020.

[29] Moni.N, Benny.P,Oblivious Polynomial Evaluation„ 31st STOC, 1999.

[30] J.Moon ,S.Jung ,S.Park , and E.Hwang. Conditional Tabular GAN-Based Two-Stage Data Generation Scheme for Short-Term Load Forecasting, November 10, 2020,

[31] `https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/`,consulté le 15/05/2022.

[32] `https://ledatascientist.com/google-colab-le-guide-ultime/`, consulté le 10/05/2022 .

[33] `https://datascientest.com/kaggle-tout-ce-quil-a-savoir-sur-cette-plateforme`, consulté le 14/05/2022.

[34] `https://datascientest.com/pandas-python-data-science`, consulté le 13/05/2022

[35] `https://www.w3schools.com/python/numpy/numpy_intro.asp`, consulté le 14/05/2022.

[36] `https://pypi.org/project/ctgan/`, consulté le 12/05/2022.

[37] `https://matplotlib.org/`, consulté le 17/05/2022.

[38] `https://scikit-learn.org/stable/`, consulté le16/05/2022.

[39] `https://pypi.org/project/table-evaluator/`, consulté le 12/05/2022

[40] Regina L. Nuzzo, PhD,Histograms : A Useful Data Analysis Visualization, American Academy of Physical Medicine and Rehabilitation, 2019.

[41] B.Liu. Web Data Mining, Exploring Hyperlinks, Contents and Usage Data. Springer, Berlin, 2011. 852

[42] Pham, Duc Truong and Dimov, Stefan S and Nguyen, Chi D, Selection of K in K-means clustering, Proceedings of the Institution of Mechanical Engineers, Part C : Journal of Mechanical Engineering Science,2005.

[43] Pham, Duc Truong and Dimov, Stefan S and Nguyen, Chi D, Selection of K in K-means clustering, Proceedings of the Institution of Mechanical Engineers, Part C : Journal of Mechanical Engineering Science,2005.

[44] C.C.Aggarwal,P.S.Yu, Privacy Preserving Clustering Data mining Model and algorithms, 2008.

[45] R.R.Rajalaxmi,A.M.Natarajan, An Effective Data Transformation Approach For Privacy Preserving Clustering, Journal of Computer Science, 320-326,2008.