République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Ain Temouchent Belhadj Bouchaib

Faculté des sciences et technologies
Département des Mathématiques et de l'Informatique

Mémoire

En vue de l'obtention du Diplôme de Master en Informatique

Option :

Réseaux et Ingénierie des Données (RID)

Thème

Anonymisation des données sensible dans un data-set.

Présenté par :
M^{lle}ATTOU Amina
M^{r} DJEDOUI Ibrahim

Devant le jury composé de :

Président :	M^{r} *SAIDI* Mohamed Reda	MCA	U.A.T.B.B.
Examinateur :	M^{r} *BENOMAR* Mohamed Lamine	MCA	U.A.T.B.B.
Encadrant :	M^{r} *BENARIBI* Imad Fethi	MCB	U.A.T.B.B.

Année universitaire 2022/2023

Dédicace

Je dédie ce modeste travail :

A celle m'a donné la vie, le symbole de tendresse, qui s'est sacrifiée pour mon bonheur et ma réussite, qui ma soutenu dans les moments les plus difficiles,
« ma chère Maman » je lui dis merci ...

A mon père, l'école de mon enfance, qui a été mon ombre durant toutes les années d'études, et qui a veillé tout long de ma vie a m'encourager, a me
donner l'aide et me protéger.

A mes chères quatre frères qui ont été toujours la pour moi, mes belles sœurs,

mes neveux et mes nièces, et a toute ma famille. Que Allah garde et protège ma famille.

A tous mes amis, mes collèges de classe, a tous les personnes qui m'ont
souhaité la réussite.

ATTOU Amina

Dédicace

Je dédie humblement ma thèse à vous tous, ceux qui ont été mes piliers de force tout au long mon parcours académique, je souhaite exprimer ma profonde gratitude et mon amour pour vous tous. Merci du fond du cœur pour votre amour, votre compréhension et votre soutien inconditionnel.

Papa, même si tu n'es plus parmi nous tu as été mom modèle de sagesse et courage. Tu m'as appris à ne jamais abandonner, peu importe les difficultés rencontrées sur mon chemin.

Maman, depuis mon enfance tu m'as inculqué la discipline et l'importance de l'éducation tu as toujours cru en moi même lorsque j'ai douté de moi même, tu as été le première enseignante de la vie, aujourd'hui je te souhaite te remercier du fond de cœur de tous ce que tu as fait pour moi.

Je vous aimes plus que les mots ne peuvent le dire

DJEDOUI Ibrahim

Remerciement

Tout d'abord on remercie Dieu de nous avoir donnée la force et la capacité de terminer ce projet modeste qui fait notre fierté et notre personnalité.

On tient a remercie toutes les personnes qui nous ont apporté leur aide et qui ont contribué a l'élaboration de ce mémoire ainsi qu'a la réussite de cette formidable année universitaire
malgré les conditions sanitaire.

On tient a remercie sincèrement notre encadreur M^r $BENARIBIIMADFATHI$, et Doyen de la Faculté des Sciences, Département des mathématiques et informatiques qui m'a tout d'abord, accueilli pour avoir encadré nos travaux de recherche je le remercie pour leur confiance qu'il m'a accordé, et pour ses précieux conseils et son suivi durant la réalisation de
notre mémoire.

Enfin on remercie tout l'équipe pédagogique de $l'universit\ d'Ain\ Tmouchent\ BelhadjBouchaib$ pour leur générosité, leur disponibilité durant toutes ces années.

Résumé

L'anonymisation des données sensibles est un processus important pour garantir la protection de la vie privée des individus dans les data-sets, il est nécessaire de recourir à des algorithmes complexes afin de rendre les données personnelles anonymes. Ces algorithmes sont conçus pour réduire au maximum le risque de réidentification tout en préservant l'utilité des données. On à travaillé avec des outils d'anonymisation principaux pour garantir la confidentialité et la sécurité des données sensibles. Enfin, nous avons testé des algorithmes d'anonymisation sur une base de donnée pour évaluer l'impact des différents paramètres de ces techniques d'anonymisation.

Mots clés : sécurité, anonymisation, données sensibles,confidentialité.

Abstract

The anonymisation of sensitive data is an important process for guaranteeing the protection of individual privacy in data-sets. It is necessary to use complex algorithms to anonymise personal data. These algorithms are designed to minimise the risk of reidentification while preserving the usefulness of the data. In our dissertation we worked with the main anonymisation tools to guarantee the confidentiality and security of sensitive data.

At last, we tested anonymisation algorithms on a database to assess the impact of the different parameters of these anonymisation techniques.

Keywords : security, anonymisation, sensitive data, privacy protection, confidentiality.

ملخص

يُعد إخفاء هوية البيانات الحساسة ،عملية مهمة جدا لضمان حماية الخصوصية الفردية في مجموعات بيانات . و من الضروري استخدام خوارزميات معقدة لإخفاء هوية البيانات الشخصية. و قد تم تصميم هذه الخوارزميات لتقليل مخاطر إعادة تحديد الهوية مع الحفاظ على فائدة البيانات.و في بحثنا هذا قد استعملنا أدوات إخفاء الهوية الرئيسية لضمان سرية وأمن البيانات الحساسة. و في الأخير، اختبرنا خوارزميات إخفاء الهوية في قاعدة بيانات لتقييم تأثير المعلمات المختلفة لتقنيات إخفاء هذه الهوية.

كلمات مفتاحية: : الحماية ، إخفاء الهوية، البيانات الحساسة ، حماية الخصوصية ، السرية .

Table des matières

Table des figures **7**

Liste des tableaux **8**

Introduction générale **9**

1 la presévation de la vie privée **10**
1.1 Introduction 11
1.2 Protection des données 11
1.3 Menaces sur la vie privée 12
1.4 Confidentialité et sécurité 12
1.5 Le chiffrement des données 13
1.6 Conclusion 14

2 Anonymisation des données sensibles **15**
2.1 Introduction 16
2.2 L'anonymisation des données 17
2.3 L'anonymat et la confidentialité des données 17
2.4 les information relatives aux données privées 19
2.5 Les techniques de base d'anonymisation des données 20
2.6 Les méthodes de la préservation de la vie privée 23
2.6.1 Algorithme de k-anonymat : 23
2.6.2 Algorithme de L-diversité : 25
2.6.3 Algorithme de la t-proximité : 27
2.6.4 Algorithme de confidentialité différentielle (differential privacy) : 28
2.7 Objectifs d'anonymisation 29
2.8 L'inconvénient de l'anonymisation 30
2.9 Conclusion 31

3 Les étapes d'anonymisation de données **32**
3.1 Introduction . . . 33
3.2 Processus de nettoyage . . . 33
3.3 Processus de généralisation . . . 34
3.4 Processus d'anonymisation . . . 36
3.4.1 Principe de l'algorithme K-Anonymat . . . 36
3.4.2 Principe de ϵ-Differential Privacy . . . 37
3.5 Comparaison entre K-Anonymity, l-diversity et ϵ-Differential . . . 40
3.6 Conclusion . . . 41

4 Discussions et résultats **42**
4.1 Introduction . . . 43
4.2 L'Enivrement de travail . . . 43
4.2.1 Langage et bibliothèques utilisées . . . 43
4.3 Réalisation . . . 44
4.3.1 Les bases de données utilisées . . . 44
4.3.2 Estimation des performances . . . 45
4.4 Résultats . . . 47
4.4.1 La base de données " adult " : . . . 47
4.4.2 La base de données « pima » . . . 51
4.5 Discussions : . . . 55
4.6 Présentation de notre d'application : . . . 57
4.7 Conclusion . . . 59

Bibliographie **63**

Table des figures

2.1 Collecte et publication des données. 16
2.2 Compromis entre l'utilité et le respect de la vie privée[16]. 18
2.3 Généralisation par hiérarchie d'une caractéristique avec le niveau d'étude d'une personne. 22
2.4 Attaque d'homogénéité. 25

3.1 Organigramme de l'algorithme de ϵ-Differential Privacy[11]. 40

4.1 Diagramme de précision et distorsion par rapport au changement de seuil k. 48
4.2 Diagramme de précision et distorsion en indemnités de l. 50
4.3 Diagramme de précision et distorsion en indemnités de seuil k. . . . 52
4.4 Diagramme de précision et distorsion en indemnités de l. 54
4.5 l'interface de l'application. 57
4.6 charger la base de données. 58
4.7 les boutons pour faire nettoyage et insert le seuil k et l. 58
4.8 Résultats finale de l'application. 59

Liste des tableaux

2.1 Ensemble de données dont l'attribut Age a été supprimé. 21
2.2 Ensemble de données avant application de la permutation 21
2.3 Ensemble de données après application de la permutation 22

4.1 La base de données « adult.csv ». 45
4.2 La base de données « pima-indians-diabetes ». 45
4.3 Résultats de test d'algorithme « k-anonymat » avec différant seuil k. 47
4.4 bdd « adult.csv » anonymisée par l'algorithme k-anonymat. 49
4.5 Résultats de test d'algorithme " l-diversity " avec différant l. 50
4.6 Résultats de test d'algorithme « k-anonymat » avec différant seuil k du bdd" pima ". 51
4.7 bdd " pima " anonymisée par l'algorithme k-anonymat. 53
4.8 Résultats de test d'algorithme « l-diversity » avec différant l du bdd " pima ". 53
4.9 bdd « adult » anonymisée par l'algorithme ϵ-Differential Privacy. . 56

Introduction générale

L'augmentation de la capacité de stockage des données a permis de stocker d'avantage d'information, ce qui peut être utile pour analyser les tendances et les modèles. Cependant, cela a également soulevé des préoccupations quant à la protection de la vie privée. Il est important de protéger les données personnelles et de s'assurer que les informations ne sont pas utilisées à des fins malveillantes c'est pour cella les organisations propriétaires de données cherchent des moyens d'anonymiser les données afin de protéger la vie privée des individus.

L'anonymisation des données sensibles dans les bases de données est une pratique courante pour protéger la vie privée et les informations personnelles des individus, cela implique généralement la suppression ou la modification de certaines information, telles que les noms, les adresses et les numéros de sécurité pour éviter que les données ne soient associées à des personnes spécifique.
L'intérêt d'étudier l'anonymisation des données sensibles dans les bases de données est importante pour comprendre les meilleurs pratiques pour protéger la vie privée des individus et prévenir la divulgation non autorisée de données personnelles. Cela peut aider les organisations à développer des politiques et des procédures pour protéger les données sensibles.

L'anonymisation des données sensibles dans une base de données peut être un processus complexe qui exige des compétences techniques suffisantes pour être effectué correctement, il existe de nombreux algorithme et techniques d'anonymisation sensibles dans une base de données, mais trouver le bon algorithme peut être un processus complexe qui nécessite une expertise technique donc il est important de comprendre que l'anonymisation des données peut être un processus délicat qui nécessite une expertise technique pour être effectué d'une manière correcte.

L'objectif de notre projet de fin d'étude est de faire une comparaison entre deux algorithmes d'anonymisation, les algorithmes sont appliqués sur une base de données pour déduire qui est l'algorithme le plus efficace pour le processus d'anonymisation.

Chapitre 1

la presévation de la vie privée

1.1 Introduction

La sécurité en général et en informatique est essentielle pour protéger les données, les systèmes informatiques, les réseaux et les utilisateurs contre les menaces et les attaques potentielles. La sécurité informatique comprend un vaste éventail de pratiques et de technologies qui protègent les systèmes et les informations sensibles contre les piratages, les intrusions, les logiciels malveillants et les violations de données.

Les téchnologies informatiques se développent à un rythme accéléré en parallèle que les méthodes de piratage et d'intrusion, ces dernières dédiés à happer des informations sensibles et personnelles d'où la nécessité d'installer des mesures de sécurité solides.

Dans de nombreux pays, il existe des lois et des règlementations qui protègent les données personnelles, comme le règlement général sur la protection des données (RGPD) de l'union européenne. Ces lois énoncent les obligations des entreprises et des organisations en manière de collecte, de stockage, de traitement et de transfert des données personnelles.

Outre le RGPD et la Loi n°78-17, DEF Algérie s'engage à respecter la présente politique de confidentialité (ci-après la « politique de confidentialité ») dans le cadre de chacun des traitements de données a caractère personnel qu'il met en œuvre [6].

Ainsi que l'anonymisation est mise en place. L'anonymisation nécessite une expertise technique et une compréhension approfondie des techniques d'anonymisation disponibles, ainsi que de leur contexte d'utilisation.

La norme ISO/TS 25237 :2008, définit l'anonymisation comme « un processus qui supprime l'association entre l'ensemble de données identifiant et le sujet des données »[17].

1.2 Protection des données

— Le respect de la confidentialité des données implique plusieurs principes fondamentaux.

— La minimisation des données exige que seules les informations nécessaires à une finalité spécifique soient collectées.

— Le consentement explicite doit être obtenu avant toute collecte ou traitement de données personnelles.

— Le principe de souveraineté des données accorde aux individus le droit d'accéder à leurs informations personnelles, de les corriger et de les supprimer.

— La transparence est également importante, car elle garantit que les utilisateurs sont informés de la collecte, de l'utilisation et du partage de leurs données personnelles.

1.3 Menaces sur la vie privée

— La vie privée est menacée par l'utilisation non autorisée ou malveillante des données collectées, et plusieurs menaces sont à prendre en compte.

— la divulgation de données personnelles est devenue facile avec l'émergence des réseaux sociaux et des services de partage de contenu. Les informations telles que les photos, vidéos, dates de naissance, préférences musicales ou culinaires, et les lieux fréquentés augmentent les risques de préjudice, de discrimination et de perte d'autonomie, représentant un danger permanent pour l'individu.

— Le vol et l'usurpation d'identité sont des crimes en forte croissance, où les criminels utilisent les données d'une personne à leur avantage.

— le profilage consiste à compiler des dossiers d'informations sur des individus afin de déduire leurs intérêts et caractéristiques. Bien que le profilage puisse être bénéfique dans les systèmes de recommandations pour proposer des produits et des services correspondant aux préférences et intérêts des clients, il devient une menace pour la vie privée si les données sont collectées de manière illégale ou utilisées à des fins malveillantes, telles que la discrimination par les prix, les publicités non sollicitées, les spams, etc [12].

1.4 Confidentialité et sécurité

Confidentialité (vie privée) : « La confidentialité peut se définir comme l'obligation éthique, professionnelle et juridique du médecin de ne pas divulguer ce qui lui est communiqué dans le cadre de la relation médecin-patient.» [18]. La confidentialité est considérée comme un droit fondamental de l'individu et est souvent protégée par la loi. Les lois sur la confidentialité peuvent varier selon les pays et les juridictions, mais elles ont généralement pour objectif de protéger la vie privée des individus contre la collecte, l'utilisation et la divulgation non

autorisées de leurs informations personnelles.[21].

La confidentialité désigne le fait de garder quelque chose ou une information privée et secret vis à vis de tout le monde sauf les gens qui sont autorisé à le voir[19] [3].

Dans une autre définition la confidentialité est le caractère réservé d'une information dont l'accès est limité seules les personnes autorisées peuvent accéder à cette information privée[19] [3].

Sécurité : La définition simple de la sécurité est le fait de protéger les informations sensibles. La sécurité est un ensemble des mesures, techniques qui assurent la préservation des informations d'une personne ou d'une organisation contre les risques tel que la divulgation, la perte, l'accès non autorisé, l'utilisation abusive[2] [9].

La préservation de la confidentialité des données se concentre sur la gestion et l'utilisation appropriée des données individuelles, y compris la mise en place de politiques pour garantir que les informations personnelles des consommateurs soient collectés, partagées et utilisées de manière responsable. En revanche, la sécurité des données vise davantage à protéger les données contre les attaques malveillantes et l'utilisation abusive de données volées à des fins lucratives. Bien que la sécurité soit essentielle pour la protection des données, elle ne suffit pas à garantir la confidentialité des données[21].

1.5 Le chiffrement des données

Le chiffrement est une technique de sécurité utilisé pour protéger les données sensibles. Le chiffrement des données consiste à convertir les données de votre entreprise d'un format compréhensible à un format codé. Il s'agit d'une mesure de sécurité. Seule une clé unique fournie au moment du cryptage peut être utilisée pour décrypter le texte ou les chiffres compromis[13].
Le chiffrement est applicable à la fois aux données en transit et aux données au repos. Pour garantir l'accès exclusif aux données de votre entreprise, vous pouvez compléter ce mécanisme de sécurisation en l'associant à des services d'authentification. Ainsi, seuls les utilisateurs autorisés seront en mesure d'accéder aux données confidentielles de votre entreprise. Voici les deux méthodes de chiffrement des données les plus utilisées :

— Le chiffrement symétrique utilise une seule clé pour encoder et décoder.

— Le chiffrement asymétrique consiste à utiliser deux clés distinctes pour encoder et décoder les données.
— Une clé privée est partagée par le programmeur et une clé publique est partagée par tous les utilisateurs.
— Les données sont chiffrées avec la clé publique, tandis que les données déchiffrées avec la clé privée[13].

Le but du chiffrement lorsque les données sont chiffrées, elle ne peut pas être utilisée pour identifier un individu spécifique ou pour accéder à des informations sensibles. Cela permet de protéger la vie privée des utilisateurs et de réduire les risques de violation de données.

1.6 Conclusion

En mettant en place des mesures solides de sécurité informatique, de protection des données et de confidentialité, il est possible de préserver la vie privée des individus et de réduire les risques de violation de données et d'atteinte à la sécurité.

Chapitre 2

Anonymisation des données sensibles

2.1 Introduction

La publication de données préserve la vie privée est un sujet crucial dans notre société actuelle. Avec l'augmentation de la collecte de données et des violations de la vie privée, il est important de trouver des moyens de protéger les informations personnelles tout en permettent aux entreprises et aux organisations d'utiliser ces données pour améliorer leurs services.

Dans cette partie on va parler des outils de la publication de données préservant la confidentialité en anglais « privacy preseving data publishing » sont des techniques qui permettent de publier des données tout en préservant la vie privée des individus dont les données sont incluses dans les ensembles de données. Ces techniques incluent notamment l'anonymisation et le chiffrement.

Le processus de publication des données préservant la vie privée comprend généralement deux phases distinctes, tel qu'illustré dans la Figure 2.1. Trois acteurs y sont impliqués : le propriétaire (ou "owner"), l'éditeur (ou "publisher") et le destinataire (ou "recipient") des données.[17].

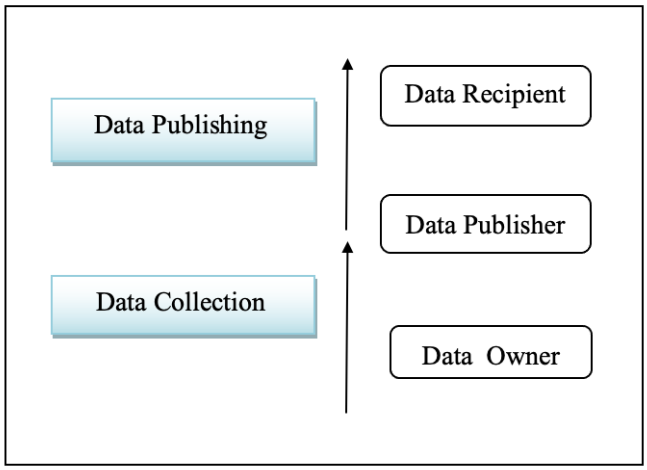

Figure 2.1: Collecte et publication des données.

2.2 L'anonymisation des données

L'anonymat signifie qu'une personne n'est pas identifiable[21]. L'anonymisation est un processus irréversible rendant impossible l'identification des personnes, toutes les informations identifiants soient directes ou indirectes sont soit modifié soit supprimé c'est à dire l'anonymisation des données sensibles consiste à supprimer ou à masquer les informations personnellement identifiables d'un ensemble de données pour protéger la vie privée des individus.[7].

La pratique de l'anonymisation des données vise à préserver la confidentialité lors de la publication de ces dernières. De nos jours, de nombreuses entreprises publiques et privées sont tenues de rendre disponibles leurs données sous forme électronique, y compris les données individuelles ou "micro-données" ; plutôt que simplement des tableaux statistiques ou des agrégats. Avant d'être publiées, ces données doivent être nettoyées en supprimant les identifiants explicites tels que les noms, les adresses et les numéros de téléphone[21].

Bien que l'anonymisation puisse sembler être une solution simple et efficace, elle comporte également des risques. Les données anonymisées peuvent encore être recoupées avec d'autres sources d'informations pour identifier les individus concernés, ce qui peut compromettre leur vie privée.

2.3 L'anonymat et la confidentialité des données

L'anonymisation des données est introduite dans le but de trouver un équilibre entre le niveau d'anonymat et le degré de confidentialité des données afin de protéger les individus contre l'utilisation abusive de leurs précieuses informations et de permettre aux chercheurs d'apprendre efficacement à partir des données. Entre l'anonymat et la perte de données[16].

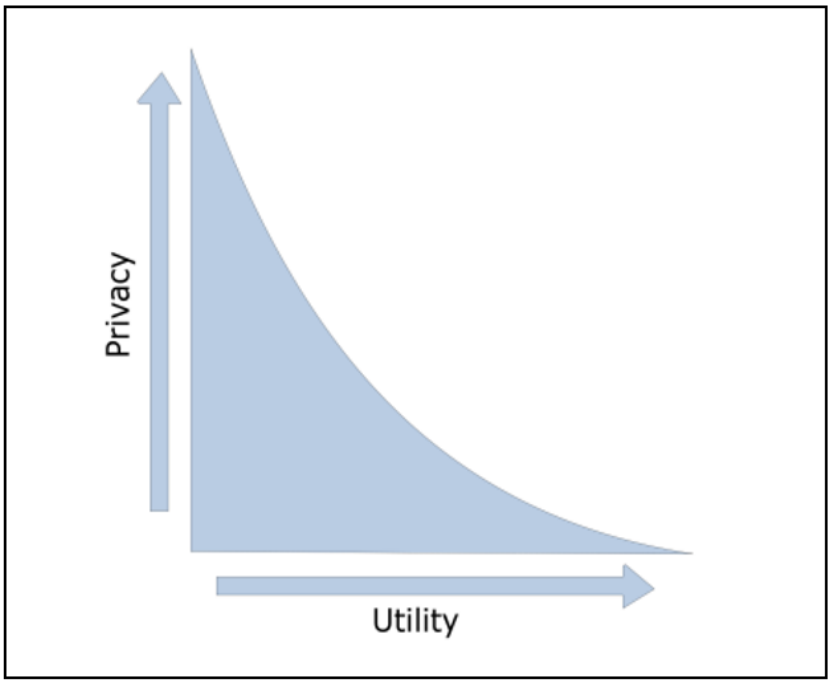

Figure 2.2: Compromis entre l'utilité et le respect de la vie privée[16].

Dans de nombreux domaines, en particulier en ce qui concerne la collecte et l'utilisation des données personnelles, il existe un compromis entre l'utilité et la confidentialité. L'utilité fait référence à la valeur ou aux avantages que les individus ou les organisations peuvent obtenir en utilisant des données. La confidentialité, quant à elle, concerne la protection des informations personnelles et la limitation de leur accès.

— Lorsque la confidentialité est élevée, cela signifie que les mesures de protection des données sont strictes, ce qui peut limiter l'accès aux données et réduire leur utilité potentielle. Cela est souvent dû à des préoccupations éthiques ou à des réglementations visant à protéger la vie privée des individus.

— D'un autre côté, lorsque l'utilité est élevée, cela signifie que les données sont largement utilisées pour générer des informations précieuses et des avantages concrets. Cependant, cela peut également entraîner des préoccupations en matière de confidentialité, car l'utilisation extensive des données peut augmenter les risques d'abus ou de violation de la vie privée.

— Il est important de trouver un équilibre entre l'utilité et la confidentialité, en mettant en place des politiques et des mesures de sécurité appropriées pour protéger les données tout en permettant une utilisation responsable et bé-

néfique. Cet équilibre dépendra du contexte spécifique et des considérations propres à chaque situation[16].

2.4 les information relatives aux données privées

Un tuple : est une structure de données qui peut contenir un ensemble de valeurs hétérogènes (c'est-à-dire de différents types de données) et qui est considérée comme un seul objet. Dans le contexte de la gestion de données, un tuple peut être considéré comme une ligne d'une table, où chaque colonne représente un attribut spécifique du tuple. Par exemple, dans une table de contacts, chaque ligne pourrait représenter un tuple avec des attributs tels que le nom, l'adresse e-mail et le numéro de téléphone. Les tuples sont utilisés dans de nombreux systèmes de gestion de bases de données pour stocker et récupérer des informations[5].

un attribut :est une caractéristique ou une propriété qui est associée à une entité. Dans le contexte des données privées, un attribut peut être une information telle que le nom, l'adresse, la date de naissance, le numéro de téléphone, le numéro du sécurité sociale, le salaire, etc. Les attributs peuvent être sensibles ou non sensibles en fonction de la nature de l'information qu'ils contiennent. Les attributs sont généralement organisés en tuples, qui sont des ensembles d'attributs qui décrivent une entité particulière dans une base de données[4].

Un identifiant explicite (IE) : fait référence à un attribut ou à un ensemble d'attributs qui permettent d'identifier directement une personne, comme par exemple un numéro de sécurité sociale, un prénom ou un nom. Bien qu'il puisse y avoir plusieurs personnes ayant le même prénom et/ou un nom ces informations nominatives peuvent mener facilement à la réidentification d'une personne au sein d'un jeu de données. Il convient de noter que l'IE ne correspond pas nécessairement à un identifiant tel que défini en modélisation conceptuelle[17].

Un quasi-identifiant (QI) :est constitué d'un groupe d'attributs qui ont la capacité d'identifier de manière indirecte au moins un individu parmi ceux figurant dans un tableau, en reliant ces attributs à des sources de données externes. Un exemple courant de quasi-identifiant est l'ensemble sexe, code postal, date de naissance, qui peut être trouvé dans de nombreuses collections de données. Ces attributs sont assez spécifiques pour permettre l'identification d'un individu unique, même dans un grand ensemble de données[17].

Un attribut sensible (AS) : les données considérées comme étant des attributs sensibles (AS) sont généralement des informations que les individus pré-

fèrent garder confidentielles, telles que des données médicales ou des informations sur les salaires[17].

Un attribut non sensible (ANS) :est un attribut qui ne contient pas d'informations personnelles ou confidentielles et qui ne peut pas être utilisé pour identifier directement ou indirectement un individu. Il s'agit d'un attribut qui n'appartient pas aux catégories d'identifiant explicite ou de quasi-identifiant[17].

2.5 Les techniques de base d'anonymisation des données

Avant d'aborder la question de l'anonymisation des données, il convient de souligner l'importance de réaliser une étape préliminaire de pseudo-anonymisation pour éliminer toutes les informations directement identifiables du jeu de données[10].

•**La pseudo-anonymisation :** La pseudo-anonymisation est une méthode courante pour anonymiser les données personnelles en supprimant les informations qui pourraient identifier une personne de manière unique, comme le nom, l'adresse, le numéro de téléphone et le numéro de sécurité sociale. Cependant, cette méthode n'est pas considérée comme fiable pour protéger les données des individus car elle peut être facilement réidentifiée en utilisant des quasi-identifiants tels que le code postal, la date de naissance et le sexe. La chercheuse Sweeney a démontré en 2001[14] que jusqu'à 80% de la population américaine pouvait être identifiée de manière unique ou presque unique en croisant deux bases de données contenant des informations nominatives et pseudo-anymisées. Ces résultats montrent que les données personnelles publiques doivent être traitées avec prudence pour éviter toute atteinte à la vie privée des individus. Malheureusement, les données de santé et autres données personnelles sont souvent accessibles au public sous cette forme, malgré les risques encourus[21].

•**Bruitage des données (Noise Data) :**L'une des techniques d'anonymisation de données les plus couramment utilisées est l'ajout de bruit, qui a été employée par des entreprises technologiques majeures telles que Google. Cette méthode consiste à altérer légèrement les attributs du jeu de données en les rendant moins précis. Par exemple, on pourrait ajouter ou soustraire quelques jours ou mois à une date. Bien que cela permette de masquer les valeurs réelles, il est important de prendre en compte le niveau de bruit nécessaire pour minimiser l'impact sur l'analyse des données et sur la vie privée des individus[21].

•**Suppression :**La technique consiste à supprimer un attribut du jeu de données lorsqu'il est inutile ou ne contribue pas à l'analyse, ou lorsqu'il est impossible de l'anonymiser autrement. Le principal avantage de cette méthode est que la suppression permanente d'un attribut ou d'un enregistrement rend impossible la récupération d'informations[21].

Cette technique crée une table anonyme où toutes les données sont dans un fichier. Les sources du tableau d'origine sont supprimées pour le risque de réidentification. Il existe deux types de suppression : suppression totale qui désigne la suppression de tuples dans leur totalité et la suppression locale qui désigne la suppression de quelque donnée de tuples[12].

	Nom	*Profession*	*Ville*	*Age*	*Sexe*
1	X	Etudiant	Alger	*	Femme
2	X	Etudiant	Oran	*	Femme
3	X	Employé	tlemcen	*	Femme
4	X	Ingénieur	Oran	*	Femme

Table 2.1: Ensemble de données dont l'attribut Age a été supprimé.

•**L'algorithme de permutation (Shuffling) :**La technique de permutation peut être utilisée sur un attribut quasi-identifiant ou sur un attribut sensible. Elle consiste à permuter les valeurs d'un même attribut au sein d'un sous-ensemble des tuples défini, comme son nom l'indique. Par exemple, dans les tableaux ci-dessus 2.2 et 2.3 on a appliqué la permutation entre le tuple 1 et 4 dans un semble de données[21].

	Nom	*Profession*	*Ville*	*Age*	*Sexe*
1	X	**E**tudiant	Alger	20	Femme
2	X	Etudiant	Oran	18	Femme
3	X	Employé	tlemcen	33	Femme
4	X	**I**ngénieur	Oran	40	Femme

Table 2.2: Ensemble de données avant application de la permutation

	Nom	*Profession*	*Ville*	*Age*	*Sexe*
1	X	**I**ngénieur	Alger	20	Femme
2	X	Etudiant	Oran	18	Femme
3	X	Employer	tlemcen	33	Femme
4	X	**E**tudiant	Oran	40	Femme

Table 2.3: Ensemble de données après application de la permutation

•**Généralisation par hiérarchie :** La généralisation consiste à délayer l'information pour qu'elle ne puisse associée à une personne ou à un petit groupe de personnes[12]. Google utilise également l'approche de généralisation par hiérarchie, qui consiste à généraliser les attributs en modifiant leur échelle ou leur ordre de grandeur. Par exemple, l'attribut "date" (jour/mois/année) peut être remplacé par l'attribut "année", en supprimant le jour et le mois. Bien que cette méthode, tout comme l'ajout de bruit, puisse empêcher l'identification de la personne, elle peut ne pas être suffisante pour assurer une anonymisation efficace[21].

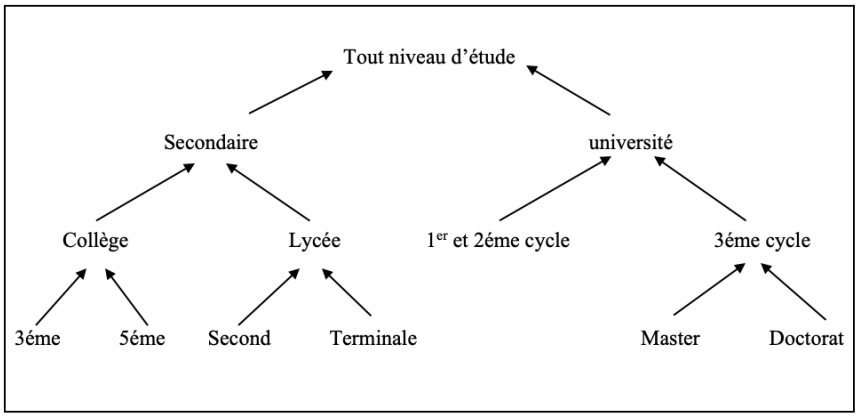

Figure 2.3: Généralisation par hiérarchie d'une caractéristique avec le niveau d'étude d'une personne.

•**Masquage (Data Masking) :**Le masquage de données (Data Masking) est une technique utilisée en conjonction avec la généralisation, qui consiste à masquer certaines valeurs dans les quasi-identifiants. Le masquage remplace la valeur supprimée par un caractère générique tel qu'un astérisque (*). Cette technique peut être appliquée aux hiérarchies de généralisation de domaine et de valeur[21].

2.6 Les méthodes de la préservation de la vie privée

2.6.1 Algorithme de k-anonymat :

est un modèle de préservation de la vie privée, le concept de k- anonymat a été inséré dans la sécurité et la confidentialité de l'information. L'idée de cette méthode est de combiner des ensembles de données avec les mêmes attributs[15]. Dans le k-anonymat la publication de données est considérée comme anonyme si les informations de chaque personne dès la publication ne peuvent être vu par au moins k-1 personne. Dans ce modèle, une base de données est une table de n lignes où chaque ligne représente un enregistrement d'une personne[12].
K-anonymat indique la possibilité de restaurer l'identification de ses enregistrements, dans l'ensemble de données si les quasi-identifiants sont identiques à au moins k-1 donc l'ensemble de donnée est considérée comme k-anonyme[1].

Algorithm 1 K-Anonymity [21]

Input: GD est l'ensemble de données généralisées;
K est le nombre de récurrence des enregistrements;
QID est la liste des Quasi-identifiers.
Output: AD ensemble des data anonymisées.

$AD \leftarrow \emptyset$ ▷ Initialiser l'ensemble des données anonymisées
$TD \leftarrow GD.groupby(QID)$ ▷ Appliquer un groupe-by à GD sur les QID
for $sousGroupe \in TD$ **do** ▷ Parcourir les sous-groupes obtenus
  if count($sousGroupe$) $>= K$ **then**
    $AD.add(sousGroupe)$ ▷ Intégration des sous-groupes dans AD
return AD

Exemple : Supposons que nous avons un ensemble de données contenant l'âge, le sexe, le code postal et le salaire annuel de 10 000 individus. Pour appliquer l'algorithme k-anonymat, nous devons garantir que chaque individu dans l'ensemble de données est indiscernable d'au moins k-1 autres individus.

— **Etape1 :**Identifier les groupes de données sensibles Dans cet exemple, les groupes de données sensibles sont l'âge, le sexe, le code postal et le salaire annuel.

— **Etape2 :**Identifier les combinaisons de groupes de données sensibles Nous devons identifier toutes les combinaisons possibles de groupes de données sensibles. Dans notre exemple, il y a 15 combinaisons possibles :
- Âge, sexe, code postal - Âge, sexe, salaire annuel - Âge, code postal, salaire annuel - Sexe, code postal, salaire annuel - Âge, sexe - Âge, code postal - Âge, salaire annuel - Sexe, code postal - Sexe, salaire annuel - Code postal, salaire annuel - Âge - Sexe - Code postal - Salaire annuel - Toutes les données.

— **Etape3 :**Grouper les données sensibles Nous devons maintenant grouper les données sensibles pour chaque combinaison de groupes de données sensibles. Par exemple, nous pouvons grouper les âges en intervalles de 10 ans, regrouper les codes postaux par région géographique, regrouper les salaires annuels par tranche de 10 000 euros, et regrouper les sexes en masculin ou féminin.

— **Etape4 :**Appliquer l'algorithme k-anonymat Pour chaque groupe de données sensibles, nous devons garantir qu'il y a au moins k-1 individus dans le groupe. Si ce n'est pas le cas, nous devons soit combiner le groupe avec un autre groupe de données sensibles, soit supprimer le groupe de données sensibles.

Par exemple, supposons que nous avons un groupe de données sensibles comprenant des individus de sexe masculin, âgés de 20 à 29 ans, vivant dans un certain code postal et ayant un salaire annuel de 35 000 euros. Si ce groupe ne compte que 4 individus, nous devons soit le combiner avec un autre groupe de données sensibles, soit le supprimer.
Une fois que nous avons appliqué l'algorithme k-anonymat à toutes les combinaisons de groupes de données sensibles, nous avons un ensemble de données protégées qui garantit que chaque individu est indiscernable d'au moins k-1 autres individus.

Pour remédier aux lacunes de la k-anonymat, d'autres techniques de regroupement ont été développées, notamment la L-diversité et la T-proximité. Ces deux méthodes améliorent l'anonymat en s'assurant que chacune des classes a des valeurs L différentes (l variation) et que les classes générées sont similaires à la distribution initiale des données[10].

•Les attaques contre (K-anonymat)

— **L'attaque d'homogénéité :** Cette attaque profite de la situation où toutes les valeurs ont une valeur sensible dans un ensemble de k enregistrements identiques, la prédiction de la valeur sensible pour l'ensemble des K enregistrements est possible même si les données ont être anonymisées[20].

Cette attaque consiste à tromper les utilisateurs pour qu'ils fournissent des informations sensibles. Par exemple, un email frauduleux peut demander à un utilisateur de fournir ses informations de connexion à un service. On a un exemple pour mieux comprendre l'attaque d'homogénéité, proposant que l'attaquant a des connaissances sur la victime comme le code postal 1825 et l'âge 28 donc l'attaquant peut conclure que le salaire de la victime est 6000.000[20].

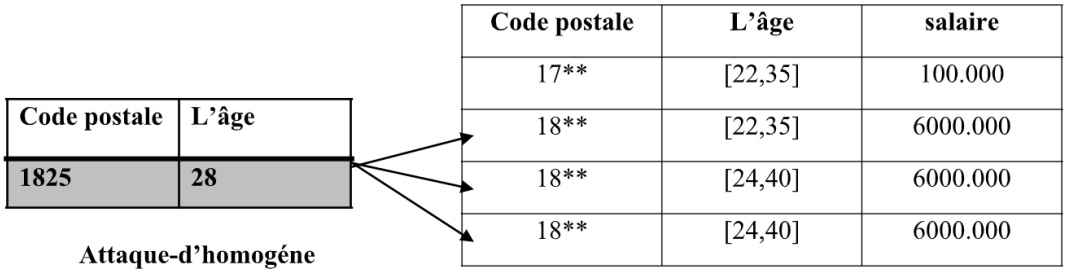

Figure 2.4: Attaque d'homogénéité.

— **Attaque par connaissance de fond :**cette attaque consiste a déduire des informations sur des individus a partir de données que ne contiennent pas directement ces informations.
Par exemple Alice connaît une personne nommée Umeko, qui a été admise dans le même hôpital que Bob. Les dossiers des patients, y compris celui d'Umeko. Alice dispose de certaines informations sur Umeko, notamment qu'elle est une jeune Japonaise de 21 ans qui habite actuellement sous le code postal 13068. En se basant sur ces données, Alice est en mesure de déduire que les informations d'Umeko sont contenues dans l'un des quatre premiers tuples du tableau.
Toutefois, sans plus de détails, Alice ne peut être sûre si Umeko souffre d'une infection virale ou d'une maladie cardiaque. Étant donné que les Japonais ont une faible incidence de maladies cardiaques, Alice en conclut avec une quasi-certitude qu'Umeko est atteinte d'une infection virale[21].

2.6.2 Algorithme de L-diversité :

Dans le modèle de L-diversité on ajoute une restriction supplémentaire sur les classes d'équivalence, Non seulement du moins k n- tuplets doivent appa-

raître dans la classe de valence, mais en plus du champ sensible associé à la classe d'équivalence doit prendre au moins des valeurs caractéristique[22].
La l-diversité (ou diversité k-anonyme) est une technique utilisée en protection de la vie privée pour préserver la confidentialité des données sensibles lorsqu'elles sont partagées ou publiées. Elle consiste à masquer ou à généraliser certaines caractéristiques des données pour réduire les risques de réidentification, un exemple d'l-diversité serait le suivant :

Algorithm 2 L-Diversity [21]

Input: GD est l'ensemble de données généralisées
L est le nombre minimal de la valeur sensible dans une class-d'équivalence
QID est la liste des Quasi-identifiers
SD est la colonne qui représente les données sensibles
Output: AD ensemble des data anonymisées.

1: $AD \leftarrow \emptyset$ ▷ Initialiser l'ensemble des data anonymisées
2: $TD \leftarrow GD.groupby(QID)$ ▷ Appliquer un groupe-by à GD sur les QID
3: **for** $sousGroupe \in TD$ **do** ▷ Parcourir les sous-groupes obtenus
4: ▷ Intégration dans AD des sous-groupes avec un nombre de valeurs diversifiées supérieur ou égal à L dans la colonne SD
5: **if** $count\_val(sousGroupe, SD) >= L$ **then**
6: $AD.add(sousGroupe)$
7: **return** AD

Exemple : Supposons que nous ayons une base de données contenant des informations sur les patients d'un hôpital, y compris leur nom, leur adresse, leur date de naissance, leur sexe, leur état de santé et leur traitement médical.
Nous voulons partager ces données avec des chercheurs pour des études statistiques, mais nous ne voulons pas révéler l'identité des patients.

Pour protéger la confidentialité des données, nous pouvons utiliser l'i-diversité en généralisant les caractéristiques sensibles.

Par exemple, nous pourrions remplacer le nom du patient par une valeur aléatoire unique, comme un identifiant de patient, pour éviter toute réidentification. Nous pourrions également généraliser l'adresse en remplaçant

l'adresse exacte par le code postal, ou en utilisant un niveau de précision différent pour les différentes caractéristiques.

De cette façon, les données restent utiles pour les études statistiques, mais il devient beaucoup plus difficile pour quelqu'un de relier les données à une personne spécifique.
En outre, pour assurer une i-diversité efficace, nous pourrions également imposer une contrainte de diversité en spécifiant que chaque groupe de k patients doit avoir au moins i valeurs distinctes pour chaque caractéristique sensible.

Par exemple, nous pourrions exiger que chaque groupe de cinq patients ait au moins deux sexes différents ou deux traitements médicaux différents pour éviter une divulgation involontaire.

•Les attaques contre l-diversité
Cette technique est vulnérable aux attaques d'asymétrie et de similarité et ne peut donc empêcher la divulgation d'attributs[21].

— **Attaque d'asymétrie :** quand la distribution générale est asymétrique la satisfaction de l-diversité n'empêche pas la détection des attributs[21].

— **Attaque par similarité :** Un attaquant peut découvrir les informations importantes, si les valeurs des attributs sensibles d'une classe d'équivalence sont particulières mais sémantiquement identique[21].

2.6.3 Algorithme de la t-proximité :

Est une méthode d'anonymisation à l'objectif de préserver la vie privée des personnes cette technique consiste perturber les données par l'ajout d'une petite quantité de bruit pour éviter que les informations ne soient pas reliées à une personne[23].
On introduit le modèle de la t-proximité pour essayer de diminuer la donnée qui peut être vu directement toujours à partir d'un regroupement de données en classe d'équivalence selon le processus du k-anonymat, cette technique est basée sur une connaissance générale de la distribution des données sensibles[22].
L'algorithme de la t-proximité permet de contrôler le degré d'anonymat en fonction du paramètre "t". Plus le paramètre "t" est élevé, plus les données seront généralisées, offrant ainsi un niveau plus élevé d'anonymat, mais au détriment de la précision des informations. En revanche, si le paramètre "t"

est faible, les données seront moins généralisées, préservant ainsi une certaine utilité, mais au risque d'une potentialité de réidentification [11]. L'objectif principal de l'algorithme de la t-proximité est de trouver un juste équilibre entre l'anonymat et la précision des données, en tenant compte des besoins spécifiques de confidentialité et d'utilité de chaque situation.

Exemple : Supposons qu'une entreprise souhaite partager des données de transaction anonymisées avec des chercheurs pour étudier les habitudes d'achat des clients, tout en préservant la confidentialité des informations personnelles. Tout en préservant les tendances globales et les informations statistiques importantes. Donc on a les données de transaction peuvent contenir des informations sur les produits achetés, le montant dépensé et la date de l'achat. Pour anonymiser ces données, l'algorithme de la t-proximité peut remplacer les produits spécifiques par des catégories générales, comme "produit A", "produit B", etc. De plus, les montants dépensés peuvent être regroupés dans des plages, comme 0-10 €, 10-20€, etc. La date de l'achat peut également être généralisée en utilisant des intervalles de temps plus larges, comme les mois ou les trimestres.

- **Attaque contre t- proximité(t-closeness)**

Défaut majeur de ce modèle est qu'il nécessite que la distribution d'un attribut sensible dans toute classe d'équivalence et la distribution d'un attribut sensible dans le tableau global soit proche, il est vraiment difficile d'identifier la proximité entre la valeur t et les connaissances acquise si en utilisant la mesure de la distance de Earth Mover en proximité[21].

2.6.4 Algorithme de confidentialité différentielle (differential privacy) :

est un moyen de protéger les données individuelles en ajoutant du bruit aux résultats. Cette technique permet de protéger la vie privée des personnes tout en fournissant des résultats globaux pertinents lorsqu'une requête est effectuée sur une base de données en ajoutant un petit nombre aléatoire aux résultats. La technique de confidentialité différenciée protège la confidentialité des données lorsqu'elles sont utilisées pour effectuer des analyses statistiques ou des calculs. Elle garantit que les résultats de ces analyses ne révèlent pas d'informations personnelles sur les personnes qui ont contribué aux données[11].

Exemple : Supposons que nous ayons une base de données contenant des informations sur le revenu des individus dans une certaine région. Nous voulons obtenir une estimation de la moyenne des revenus dans cette région, mais nous voulons également garantir la confidentialité des individus.

-Sans ϵ-Differential Privacy : Sans utiliser la technique ϵ-Differential Privacy, nous pourrions simplement calculer la moyenne des revenus en utilisant l'ensemble des données disponibles. Cependant, cela pourrait potentiellement révéler des informations sur des individus spécifiques si leurs revenus sont extrêmement élevés ou bas.

Avec ϵ-Differential Privacy : En utilisant la technique ϵ-Differential Privacy, nous introduisons du bruit aléatoire dans le calcul de la moyenne des revenus. Supposons que nous fixions un paramètre de confidentialité ϵ, qui détermine le niveau de protection de la vie privée que nous voulons garantir. Plus ϵ est petit, plus la confidentialité est renforcée, mais cela peut entraîner une perte de précision dans les résultats. Pour calculer la moyenne des revenus avec ϵ-Differential Privacy, nous ajoutons du bruit aléatoire au calcul. Par exemple, nous pourrions ajouter un nombre aléatoire généré à partir d'une distribution laplacienne au résultat de la moyenne. Ce bruit aléatoire masque les contributions individuelles tout en préservant les propriétés statistiques globales des données.

Lorsque nous obtenons le résultat final, nous pouvons garantir qu'il est différentiellement privé avec un niveau de confiance spécifié par ϵ. Cela signifie que même si un attaquant connaît toutes les données, il ne peut pas déterminer avec certitude les revenus individuels.

Par exemple, supposons que la véritable moyenne des revenus soit de 50 000 $, mais après l'application de la technique ϵ-Differential Privacy avec un ϵ spécifique, le résultat final est de 51 000 $. Cela signifie que l'attaquant ne peut pas dire avec certitude si un individu particulier a un revenu inférieur ou supérieur à 51 000 $, ce qui préserve la confidentialité des données.

En résumé, la technique ϵ-Differential Privacy permet de garantir la confidentialité des données en introduisant du bruit aléatoire dans les résultats des analyses statistiques, ce qui empêche la divulgation d'informations sensibles sur les individus.

2.7 Objectifs d'anonymisation

— L'objectif principal de l'anonymisation est de protéger la vie privée des personnes dont les données sont collectées et utilisées.

— L'anonymisation vise à rendre les données personnelles irréversiblement anonymes, de sorte qu'elles ne puissent plus être associées à une personne identifiée ou identifiable.

— l'anonymisation permet de minimiser les risques d'utilisation abusive des données personnelles, tels que la discrimination, la stigmatisation ou l'atteinte à la réputation.

— Peut être utilisée dans divers contextes, tels que la recherche médicale, les études de marché, la surveillance de la santé publique, la lutte contre la fraude ou le terrorisme, et d'autres domaines où la collecte de données personnelles est nécessaire pour réaliser un objectif spécifique.

— Permet de garantir que ces données peuvent être utilisées de manière responsable et éthique, sans porter atteinte à la vie privée ou aux droits fondamentaux des personnes concernées.

En résumé, l'objectif de l'anonymisation est de permettre l'utilisation de données personnelles tout en protégeant la vie privée et les droits des personnes concernées. Cela permet de concilier les objectifs de collecte de données avec les préoccupations éthiques et juridiques liées à la protection des données personnelles[17].

2.8 L'inconvénient de l'anonymisation

L'anonymisation peut présenter plusieurs inconvénients potentiels, selon la manière dont elle est mise en œuvre et utilisée. Voici quelques-uns des principaux inconvénients de l'anonymisation :

— **Perte de données :**lorsque les données sont anonymisées, certaines informations sensibles peuvent être supprimées ou masquées. Cela peut entrainer une perte de données importante, qui peut limiter l'utilité des données pour certaines applications.

— **Risque de réidentification :** même si les données ont été anonymisées, il peut toujours être possible de les réidentification en les combinant avec d'autres données. Les risques de réidentification peuvent augmenter si les données anonymisations sont publiques ou si elles sont stockées sur des serveurs en ligne.

— **Couts et complexité :**l'anonymisation peut être couteuse et complexe à mettre en œuvre, en particulier pour les grandes bases de données ou

les données très sensibles. Cela peut limiter la disponibilité des données anonymisées pour certains types d'utilisateurs.

— **Biais potentiels :**l'anonymisation peut introduire des biais potentiels dans les données, en particulier si les critères de suppression de données sont choisis de manière arbitraire ou si les données anonymisées ne représentant pas fidèlement la population d'origine.

On sait bien que l'anonymisation puisse être un moyen utile de protéger la vie privée et de garantir la confidentialité des données, elle peut également présenter des inconvénients qui doivent être pris en compte lors de sa mise en œuvre et de son utilisation[17].

2.9 Conclusion

En résumé, l'anonymisation peut aider a protéger la vie privée des individus, amélioré la sécurité des données, respecter les réglementations sur la protection des données et facilite la recherche et de garantir la confidentialité, il est aussi important de prendre en compte ces modèles d'attaque lors de la collecte et du stockage des données sensibles et de mettre en place des mesures de sécurité appropriées pour prévenir ces attaques.
Cela peut inclure des mesures de chiffrement, d'anonymisation et de contrôle d'accès aux données.

Chapitre 3

Les étapes d'anonymisation de données

3.1 Introduction

Les outils d'anonymisation sont des logiciels ou des techniques qui permettent de masquer ou de rendre anonymes les données personnelles dans un ensemble de données. Les méthodes d'anonymisation sont fréquemment utilisées pour protéger la vie privée des personnes tout en permettant l'utilisation des données à des objectives de recherche, d'analyse ou de développement de logiciels. Dans ce chapitre on a utilisé plusieurs types d'outils d'anonymisation.

3.2 Processus de nettoyage

Le but du processus de nettoyage des données est de préparer les données pour une utilisation ou une analyse future. Il vise à supprimer les erreurs, les valeurs aberrantes, les doublons, les données manquantes ou tout autre élément indésirable des données. Le processus de nettoyage garantit que les données sont cohérentes, précises et fiables, ce qui facilite les analyses et les prises de décisions basées sur ces données. Il contribue également à améliorer la qualité globale des données, ce qui est essentiel pour obtenir des résultats précis et pertinents lors de l'exploration ou de l'analyse des données[11].

Algorithm 3 nettoyage de base de données [21]

Input: *OriginalData* est l'ensemble de données brutes entrées par l'utilisateur.
Output: *OD* ensemble des données nettoyées.

1: $OD \leftarrow OriginalData$
2: **for** $attribut \in OD.columns$ **do** ▷ Parcours de tous les attributs de OD
3: **if** $attribut.containsOnlyNullData()$ **then**
4: $OD.remove(attribut)$ ▷ Suppression des attributs contenant uniquement des valeurs nulles
5: **for** $tuple \in OD$ **do** ▷ Parcours de tous les tuples de OD
6: **if** $tuple.containsNullData()$ **then**
7: $OD.remove(tuple)$ ▷ Suppression des tuples contenant des valeurs nulles
8: **return** OD

3.3 Processus de généralisation

Les modèles d'anonymisation reposent sur le processus d'anonymisation, cette étape est importante car elle consiste à modifier ou à supprimer délibérément certaines données afin de rendre moins identifiables. Les données peuvent être modifiées en créant un ensemble de plages ou en délimitant une zone étendue avec des limites appropriées. Le but est de diminuer de la précision des valeurs des attributs quasi-identifiers sans pour autant diminuer la qualité des données résultantes. À la fin de ce processus, nous obtiendrons un ensemble de données généralisées, la donnée sensible ainsi que la liste des attributs quasi-identifiers avec leurs types de données[21].

• **Définition de données sensibles :** L'utilisateur de notre système doit identifier les informations sensibles qu'il souhaite protéger au niveau de la première étape du processus de généralisation. En conséquence, nous listerons chaque caractéristique en fonction du type de données qu'elle contient. La donnée sensible est la première à être choisie car elle ne doit pas être altérée par les ajustements effectués dans les étapes suivantes[21].

•**Définition des quasi-identifiers :** La deuxième étape de la généralisation consiste à reformuler le texte. Nous faisons de nouveau appel à l'utilisateur de notre système durant cette étape afin qu'il puisse identifier les attributs quasi-identifiants de son ensemble de données[21].

•**Définition de types de généralisation :** Dans cette étape, nous demanderons à l'utilisateur de choisir un type de généralisation géré par notre système pour chaque attribut (sauf l'attribut sensible). Les types de généralisation sont masquage consiste a cachée une partie d'un attribut « U****** », suppression consiste à remplacer les valeurs d'attributs par « * », intervalle remplacer les valeurs d'un attribut par des espaces métriques définis par l'utilisateur « [0,50] ».

Algorithm 4 Généralisation de données[21]

Input: OD est l'ensemble de données originales
QID est la liste des Quasi-identifiers
$GQID$ est la liste de généralisation pour chaque Quasi-identifier
Output: GD est l'ensemble de données généralisées

$GD \leftarrow OD$ ▷ Initialiser l'ensemble des données généralisées avec OD
for $i \in [0...count(QID)]$ **do** ▷ Parcourir l'ensemble des quasi-identifiers
    ▷ Vérification du type de généralisation pour l'attribut quasi-identifiers i.
    if $GQID[i].type = "Interval"$ **then**
        for $val \in OD[QID[i]]$ **do** ▷ Parcours des valeurs de l'attribut $QID[i]$
            for $j \in [0...nombreIntervalPossible]$ **do**
                ▷ Vérification de chaque intervalle possible
                if $val < (j * GQID[i].val)$ & $val >= ((j+1) * GQID[i].val)$ **then**
                    $val \leftarrow [((j-1) * GQID[i].val), (j * GQID[i].val)[$
    else
        if $GQID[i].type = "Masquage"$ **then**
            for $val \in OD[QID[i]]$ **do**
                ▷ Parcourir et remplacer les j derniers caractères de la val
                for $j \in [(len(val) - GQID[i].val)...len(val)]$ **do**
                    $val[j] \leftarrow *$
        else
            if $GQID[i].type = "Suppression"$ **then**
                for $val \in OD[QID[i]]$ **do**
                    ▷ Remplacer chaque val par $*$
                    for $val \in OD[QID[i]]$ **do** $val \leftarrow *$
            else
                if $GQID[i].type = "Hierarchie"$ **then**
                    ▷ Remplacer chaque valeur par sa valeur hiérarchique
                    for $val \in OD[QID[i]]$ **do** $val \leftarrow GQID[i].val$
return GD

3.4 Processus d'anonymisation

Âprès avoir une généralisation sur la base de données, cette dernière est prete pour les étapes suivante pour faire l'anonymisation de données.

3.4.1 Principe de l'algorithme K-Anonymat

Le principe de K-Anonymat est une méthode de protection de la vie privée des individus dans les ensembles de données publiques ou partagées. Il vise à garantir que les données individuelles ne peuvent pas être liées à une personne spécifique en les rendant indiscernables parmi un groupe d'au moins K-1 autres individus.
Le principe de K-Anonymat repose sur la généralisation de certaines informations d'identification dans les enregistrements de données. Cela peut inclure des attributs tels que l'âge, le sexe, la localisation géographique, les revenus, etc.
L'objectif est de rendre chaque individu du groupe d'anonymat indiscernable des autres en remplaçant les valeurs spécifiques par des catégories plus générales.
Par exemple, si K est défini comme 3, alors chaque enregistrement de données doit être rendu indiscernable en le combinant avec au moins deux autres enregistrements similaires. Cela peut être réalisé en généralisant l'âge à des tranches d'âge, en remplaçant les valeurs exactes des revenus par des intervalles de revenus, ou en masquant partiellement l'adresse en ne conservant que la ville ou la région.
L'objectif ultime de K-Anonymat est d'empêcher la réidentification des individus à partir des données publiées, même en combinant ces données avec d'autres sources d'informations disponibles.
En garantissant que chaque individu est indiscernable parmi un groupe suffisamment large, il devient plus difficile de relier les données à une personne spécifique.
Cependant, il convient de noter que le principe de K-Anonymat ne fournit pas une garantie absolue de confidentialité. Des techniques avancées telles que l'attaque de réidentification basée sur les connaissances (known-plaintext attack) ou l'inférence statistique peuvent toujours potentiellement compromettre l'anonymat des individus. Par conséquent, des précautions supplémentaires et des techniques complémentaires peuvent être nécessaires pour renforcer la confidentialité des données[11].

3.4.2 Principe de ϵ-Differential Privacy

Le principe de ϵ-Differential Privacy est une approche mathématique pour protéger la vie privée des individus lors de la collecte, de l'analyse et de la publication de données. Il vise à garantir que la présence ou l'absence des données d'un individu particulier ne modifie pas de manière significative les résultats ou les analyses effectuées sur les données globales. Le principe de ϵ-Differential Privacy repose sur l'ajout de bruit contrôlé aux données afin de masquer les informations sensibles et d'empêcher la réidentification des individus.
Le paramètre ϵ est utilisé pour quantifier le niveau de confidentialité et détermine la quantité maximale de différence autorisée entre les résultats obtenus avec ou sans les données d'un individu. L'idée centrale de l'ϵ-Differential Privacy est d'introduire du bruit aléatoire dans les calculs effectués sur les données, tout en maintenant la validité statistique et l'utilité des résultats.
Le bruit est calibré de manière à respecter la valeur ϵ, qui représente le niveau de confidentialité souhaité. Plus la valeur ϵ est petite, plus la garantie de confidentialité n'est forte, mais cela peut entraîner une perte d'utilité des données. L'ϵ-Differential Privacy permet de fournir des garanties formelles et quantifiables sur la confidentialité des données. Il offre une approche robuste pour analyser les données tout en protégeant la vie privée des individus.
Cette approche est largement utilisée dans les domaines de la recherche, de l'apprentissage automatique et de l'analyse de données sensibles, tels que les données médicales ou les données de localisation. Il convient de noter que l'ϵ-Differential Privacy ne garantit pas une confidentialité absolue, mais fournit une protection probabiliste contre les attaques de réidentification et les fuites d'informations. Il est important de choisir et de calibrer
soigneusement les mécanismes de perturbation pour atteindre le bon équilibre entre la confidentialité et l'utilité des données dans chaque scénario d'application spécifique[11].

— **Mécanisme de Laplace :** La distribution de Laplace est préférée en raison de sa simplicité et familiarité, de ces propriétés statistiques, de sa preuve mathématique de confidentialité et de sa compatibilité avec les opérations statistiques. Elle offre une protection de la vie privée en introduisant un niveau de bruit élevé[11].

  Algorithme de Laplace est un mécanisme couramment utilisé en confidentialité différentielle pour ajouter du bruit aléatoire aux résultats d'une requête statistique. Le déroulement du l'algorithme de Laplace :

  — Définir la sensibilité de la fonction de requête (sensibilité).

- Définir le paramètre de confidentialité différentielle (epsilon).
- Générer un nombre aléatoire selon la distribution de Laplace avec une moyenne de 0 et une échelle basée sur la sensibilité et l'epsilon.
- Ajouter ce nombre aléatoire à la réponse de la requête.
- Renvoyer le répons bruité[11].

Soit $\epsilon \in \mathbb{R}^+$, un mécanisme d'anonymisation M, et Im_M l'image de M. M est dit ϵ-différentiel si pour tout D et D' deux jeux de données adjacents et pour chaque $D^* \in \mathrm{Im}_M$:

$$P(M(D) = D^*) < exp(\epsilon)P(M(D') = D^*) \tag{3.1}$$

La confidentialité différentielle est adaptée au cas où des requêtes sont effectuées sur le jeu de données personnelles. Dans ce cas, une méthode classique pour respecter la confidentialité différentielle est d'utiliser un mécanisme de Laplace[24].

- Calcul de la l1-sensibilité : contribution de l'individu le plus influent sur la requête f (qui peut être une moyenne, une m´médiane, etc.)[24] :

  $$\Delta_1 f = max_{D,D'}||f(D) - f(D')||_1 \tag{3.2}$$

  avec D, D'deux jeux adjacents.

- Pour une requête f, le mécanisme de Laplace défini par

  $$M(D, f, \epsilon) = f(D) + (Y_1, ..., Y_K) \tag{3.3}$$

  où (Y_i) variables aléatoires i.i.d. de loi de Laplace $Lap(\Delta 1 f/\epsilon)$ est ϵ-différentiel[24].
  Soit $(\epsilon, \delta) \in (0, 1)^2$, un mécanisme d'anonymisation M, et Im_M l'image de M. M est dit (ϵ, δ)-différentiel si pour tout D et D' deux jeux de données adjacents et pour chaque $D^* \in \mathrm{Im}_M$:

  $$P(M(D) = D^*) < exp(\epsilon)P(M(D') = D^*) + \delta \tag{3.4}$$

  Une méthode classique pour respecter la confidentialité est d'utiliser un mécanisme Gaussien[24].
- Calcul de la l2-sensibilité : contribution de l'individu le plus influent sur la requête f (qui peut être une moyenne, une m´ediane, etc.) :

  $$\Delta_2 f = max_{D,D'}||f(D) - f(D')||_2 \tag{3.5}$$

  avec D, D'deux jeux adjacents[24].

— Pour une requête f, le mécanisme Gaussien défini par

$$M(D, f, \epsilon, \delta) = f(D) + (Y_1, ..., Y_K), \tag{3.6}$$

où (Y_i) variables aléatoires i.i.d. de loi Normale $N(0, \Delta_2^2 f \frac{ln(1,23/\delta}{\epsilon^2})$ est (ϵ, δ)-différentiel [24]

Déroulement d'algorithme de confidentialité ϵ-différentiel :

— Prétraitez les données non numériques pour les affecter à des classes numériques afin de former le modèle.

— Pour chaque colonne sensible :

  — Attribuez aux valeurs cibles de classification les mêmes que les valeurs d'origine, effectuez un apprentissage supervisé et attendez-vous à ce que les résultats de prédiction soient les mêmes.

— Initialiser le modèle de classificateur GaussianNB avec ϵ et sensibilité.

— Pour chaque enregistrement du jeu de données :

  — Ajustez l'enregistrement au classificateur GaussianNB.

  — Mettez à jour la moyenne gaussienne et la variance avec le bruit laplacien ajouté à la valeur.

— Pour chaque enregistrement du jeu de données :

  — Pour chaque colonne sensible :

    — Prédisez la valeur des attributs sensibles dans l'enregistrement avec classification.

    — Écraser les données d'origine.

  — Renvoyez l'ensemble des données anonymisé[11].

Le diagramme de flux de ϵ-Differential Privacy est illustré ci-dessous

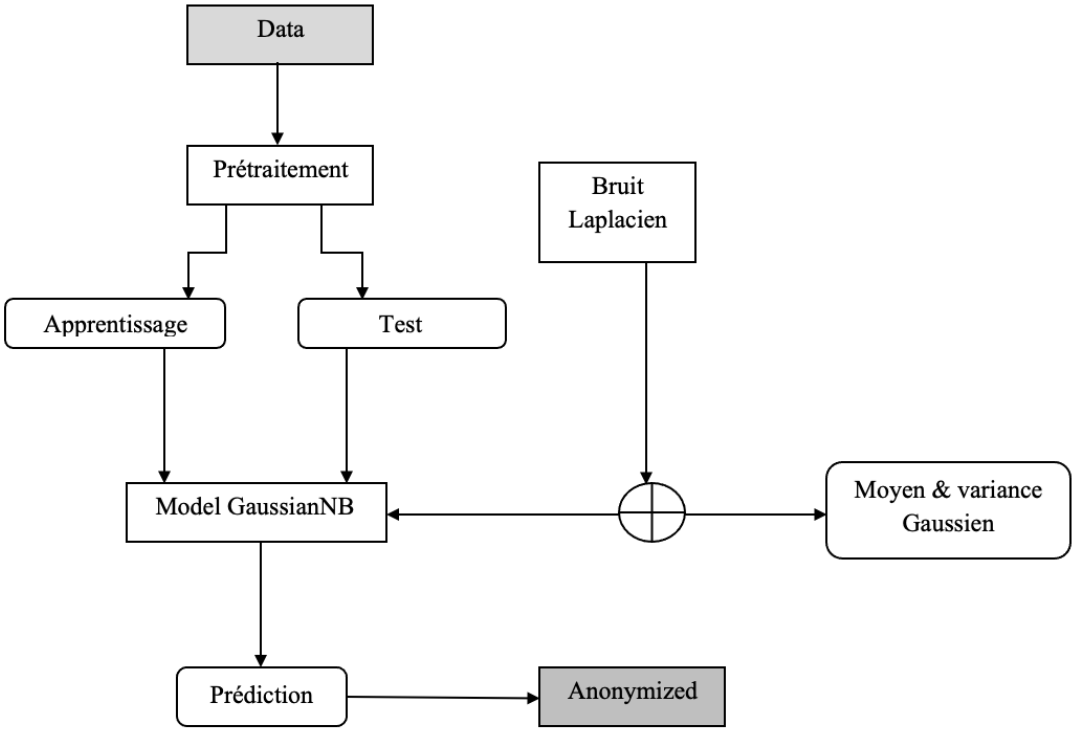

Figure 3.1: Organigramme de l'algorithme de ϵ-Differential Privacy[11].

3.5 Comparaison entre K-Anonymity, l-diversity et ϵ-Differential

K-Anonymity, l-diversity et ϵ-Differential Privacy sont deux approches différentes pour protéger la vie privée des individus lors de la divulgation ou du partage de données. Bien qu'ils visent tous deux à préserver la vie privée, ils diffèrent par leurs techniques et leurs objectifs.
Alors K-Anonymity, l-diversity se concentre sur la prévention de la réidentification en regroupant les individus, tandis que ϵ-Differential Privacy fournit une mesure quantifiable de la vie privée en ajoutant un bruit contrôlé à l'aide des fonction matématiques . Le choix entre ces techniques dépend des exigences spécifiques du scénario de partage de données, du niveau de confidentialité souhaité et de l'impact acceptable sur l'utilité des données.

3.6 Conclusion

Dans ce chapitre nous avons présenté les étapes d'anonymisation telque le processus de nettoyage, généralisation et d'anonymisation, dans le but de préserver la confidentialité pour le jeu de données (data.set).

Dans le prochain chapitre, nous allons présenter l'implémentation à l'aide des différents outils définit dans ce chapitre.

Chapitre 4

Discussions et résultats

4.1 Introduction

Après avoir décrit on détail les algorithmes k-anonymat, l-diversity et ϵ-Differential Privacy, dans ce chapitre nous allons tester ces dernier et présenter les résultats obtenues de notre travaille.

4.2 L'Enivrement de travail

4.2.1 Langage et bibliothèques utilisées

— **Python :** Python est un langage de programmation interprété qui fonctionne sur plusieurs plateformes et utilise plusieurs paradigmes. La programmation impérative structurée, fonctionnelle et orientée objet est favorisée par cela. Il possède un typage dynamique efficace, une gestion automatique de la mémoire à l'aide de raccourcis et un système de gestion d'exceptions, tout comme Perl, Ruby, Scheme, Smalltalk et Tcl. Certains pédagogues l'aiment parce qu'il est un langage facile à apprendre les concepts de base de la programmation grâce à sa syntaxe distincte des mécanismes de base[8].

— **Pandas** : est une bibliothèque de programmation Python qui permet de manipuler et d'analyser des données. En particulier, elle fournit des structures de données et des fonctionnalités de manipulation de tableaux numériques et de séries temporelles. Pandas est un programme gratuit disponible sous licence BSD2. Il tire son nom du terme Panel Data, qui signifie en français "données de panel", un concept d'économétrie pour les jeux de données qui incluent des observations pour les mêmes personnes sur plusieurs périodes de temps. Son nom fait également référence à l'expression "Analyse des données Python"[8].

— **NumPy** : La bibliothèque NumPy est une extension du langage de programmation Python qui ajoute les fonctionnalités pour effectuer des calculs numériques efficaces et manipuler de manière pratique des tableaux multidimensionnels.

Un ensemble d'objets de tableau de données appelés "ndarray"(array N-dimensionnel) est fourni par NumPy et est utilisé pour stocker et manipuler les données numériques. Les tableaux NumPy, car ils sont optimisés pour des opérations vectorisées, sont plus efficaces que les listes Python conventionnelles pour effectuer des opérations mathématiques et statistiques.

— **pyCANON** : une bibliothèque Python et une interface de ligne de commande (CLI) qui permettent d'évaluer le niveau d'anonymat d'un
ensemble de données en utilisant des techniques courantes d'anonymisation, telles que le k- anonymat, le (α,k)-anonymat, la ℓ-
diversité, l'entropie ℓ-diversité, la récursivité (c,ℓ)-diversité, la ressemblance β de base, la ressemblance β améliorée, la t-proximité et la confidentialité γ-divulgation.

— **Diffprivlib** : est une bibliothèque open-source pour le langage de programmation Python qui propose différentes fonctionnalités de confidentialité. Elle vise à fournir des outils et des méthodes pour protéger la confidentialité des données lors de l'analyse et de la publication de données sensibles. La confidentialité différentielle protège la vie privée des personnes dans un ensemble de données. Elle garantit que les résultats de l'analyse ne permettent pas de déduire des informations sensibles ou personnelles sur des individus particuliers, même si les attaquants ont des connaissances supplémentaires sur les autres individus dans l'ensemble de données.

4.3 Réalisation

Afin d'étudier les performances des algorithmes K anonymat, l diversité et ϵ différencie nous avons testé ces dernier sur deux base de donnée, et pour montrer l'efficacité de ces algorithmes, nous avons réalisé des comparaisons en utilisant des critères à savoir la précision, taux d'échec et distorsion.

4.3.1 Les bases de données utilisées

— **Adult :**est une base de données de taille [32561 lignes x 15 colonnes] qui contient des attributs (âge, Work class, fnlwgt, Hours-per-week, Native-country, Class, ...).

	Age	Work class	fnlwgt	...	Hours-per-week	Country	Class
0	60	State-gov	77516	...	40	USA	<=50k
1	50	Self-emp-not-inc	83311	...	13	USA	<=50k
2	38	Private	215646	...	40	USA	<=50k
3	53	Private	234721	...	40	USA	<=50k
4	28	Private	338409	...	40	Cuba	<=50k
...	...	. ....	....	. ...	...	...	...
32556	27	Private	257302	...	38	USA	<=50k
32557	40	Private	154374	...	40	USA	>50k
32558	58	Private	151910	...	40	USA	<=50k
32559	22	Private	201490	...	20	USA	<=50k
32560	52	Self-emp-inc	287927	...	40	USA	>50k

Table 4.1: La base de données « adult.csv ».

— **Pima-indians-diabetes :**est une base de données de taille [769 lignes x 9 colonnes] qui désigne les résultats des analyses de diabètes pour les femmes en saintes, qui contient les attributs suivant (Pregnancy, Glucose, BP, Triceps Th, Insulin, BMI, DiabtesP, age, Class).

Pregnancy	Glucose	BP	...	Insulin	MBI	Diabets	Age	Class
6	148	72	...	0	33.60	0.63	50	YES
1	85	66	...	0	26.60	0.35	31	NO
8	183	64	...	0	23.30	0.67	32	YES
1	89	66	...	94	28.10	0.17	21	NO
0	137	40	...	168	43.10	2.29	33	YES
...	...	...	...	...	...	...	...	...

Table 4.2: La base de données « pima-indians-diabetes ».

4.3.2 Estimation des performances

— **Taux d'échec (Hidden failure) :**Le rapport entre les modèles sensibles qui n'ont pas été masqués avec la méthode de préservation de la vie privée et les modèles sensibles trouvés dans les données d'origine[11].
Lors de la comparaison des algorithmes k-anonymat et ϵ-Différential, il est important de tenir compte de leur sensibilité aux échecs cachés. Un échec caché se produit lorsque des informations sensibles peuvent être déduites ou inférées à partir de données anonymisées, même si elles semblent être protégées.

Algorithm 5 Hidden Failure [21]

Input : len (DN) est la longueur de données nettoyées

len (DS) est la longueur des données après la méthode de préservation de la vie privée

Output m1

1 : m1 ←0 ; soit m1 le nombre de ligne non-masqué initialisé a 0

m2 ←0 ; soit m2 le nombre de ligne masqué initialisé a 0

2 : **pour** i ← 0 à len(DS) faire

3 : **pour** j ← 0 à len(DN) faire

4 : **si** Similaire () alors

m 1 ← m1+1 ;

5 : **sinon**

m 2 ← m2+1 ;

6 : **return** (m1, m1/len(DS))

— **Taux de précision :**signifie que les données sont toujours suffisamment précises pour être utiles, mais anonymes d'une manière qui protège la vie privée des personnes. Elle est calculée comme suit via la formule suivante :

$$Prcision = PR = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.1}$$

  — VP : c'est vrai positif.
  — VN : c'est vrai négatif.
  — FP : c'est faux positif.
  — FN : c'est faux négatif.

— **Taux distorsion :**La distorsion est une fonction mathématique qui prend les données d'origine en entrée et renvoie des données modifiées en sortie. Cette fonction peut être utilisée pour ajouter du bruit ou pour modifier les valeurs des données, tout en préservant certaines propriétés statistiques des données d'origine. Les techniques de distorsion peuvent inclure la perturbation aléatoire, la généralisation, l'échantillonnage et la suppression de données.

4.4 Résultats

On a pris une base de données nommée « aduls.csv » qui contient des différents attributs age, work class, education_num, capitale_gain... ect. On va appliquer deux algorithmes d'anonymisation (k-anonymat, et ϵ-Differential Privacy) sur la base de donnée.

4.4.1 La base de données " adult " :

— **Pour algorithme k-anonymat :**On a appliqué l'anonymisation avec l'algorithme k-anonymat sur la bdd « adult », avec plusieurs seuils k.
On obtient également des informations concernant chaque partition, tel que le taux de perte, le taux de différente précision, le taux de différente distorsion et taux d'échec. Ces informations sont présentées sous forme de tableaux ci-dessus :

Le seuil k	[2-9]	10	[15-50]	[80-100]
Précision	0.5	0.42	0.35	0.36
Distorsion	0.42	0.58	0.65	0.65
Taux d'échec	0.0	0.0	0.04	0.23
taille de bdd	30163	30163	30150	30094

Table 4.3: Résultats de test d'algorithme « k-anonymat » avec différant seuil k.

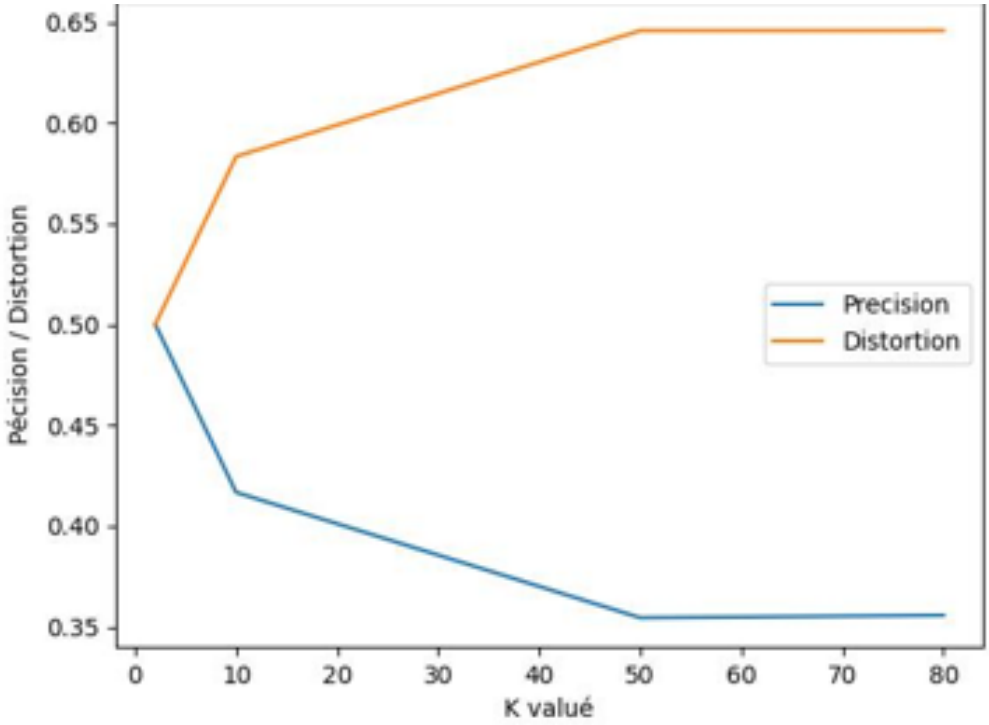

Figure 4.1: Diagramme de précision et distorsion par rapport au changement de seuil k.

-On observe que la précision diminue, la distorsion augmente à mesure que le seuil k augmente.

Observation :

- **Pour plusieur seuil k :**Cette section présente les résultats de l'anonymisation avec un seuil de k, ce qui signifie que chaque groupe d'individus doit avoir au moins [2, 10, 50, 80] membres pour garantir l'anonymat. On remarque qu'une valeur k plus élevée (par exemple, k=80) on obtient la taille du bdd moins que la taille du bdd originale, mais elle offre un niveau d'anonymat plus élevé, car elle nécessite des groupes plus grands et donc une plus grande agrégation des données, possible d'avoir une perte de donnée.
- **Pour la précision :**
  - **Pour k=[2-9] :**la précision est de 0.5. Cela signifie que les données anonymisées conservent 50% des caractéristiques présentes dans les données originales.
  - **Pour k=10 :** la précision est de 0.4167. Cela indique que les données anonymisées sont précises a hauteur de 41.67% par rapport aux les données originales.

— **Pour k=[15,80] :** on obtient la même valeur de précision 0.3544, Cela suggère que les données anonymisées ont conservé 35.44% de précision des données originales. On déduit que plus le seuil k est élevé, la précision réduit.

— **Pour la distorsion :** On remarque plus le seuil k augmente, on voit une augmentation dans la distorsion. Pour k=2, la distorsion est de 0.5 Tendit que pour k=80, la distorsion est de 0.65.

— **Pour le taux d'échec :** dans k=2 , k=10 on obtient 0% du taux d'échec, par contre plus le seuil k augmente on voit une augmentation du taux d'échec.

Exemple : pour k=50.

	Age	Work class	Education_num	...	Country	Class
0	50-100	State-gov	*	...	U***	<=50k
1	50-100	Self-emp-not-inc	*	...	U***	<=50k
2	0-50	Private	*	...	U***	<=50k
3	50-100	Private	*	...	U***	<=50k
4	0-50	Private	*	...	C***	<=50k
...	...	...	...	...	...	...
30144	0-50	Private	*	...	U***	<=50k
30145	0-50	Private	*	...	U***	>50k
30146	50-100	Private	*	...	U***	<50k
30147	0-50	Private	*	...	U***	<=50k
30148	50-100	Self-emp-inc	*	...	U***	<=50k

Table 4.4: bdd « adult.csv » anonymisée par l'algorithme k-anonymat.

On remarque que la taille de bdd a été réduite car le tuple "âge " a été regrouper dans un intervalle, par exemple la ligne 4 " l'âge " est regroupé entre [0-50], les valeurs de " education_num " a été remplacé par "*" et on a masquée juste une parie d'attribut " Native-country ".

— **Pour algorithme l-diversity :** toujours avec la base de données " adult ". On appliqué l'anonymisation avec l'algorithme l-diversity sur la bdd, avec plusieurs l.

L	[2-9]	[10-40]	[50-70]	80
Précision	99.69	96.54	69.23	50.06
Distorsion	0.31	3.46	30.47	49.94
Taux d'échec	0.31	3.46	30.77	49.94
taille de bdd	30071	29118	20883	15100

Table 4.5: Résultats de test d'algorithme " l-diversity " avec différant l.

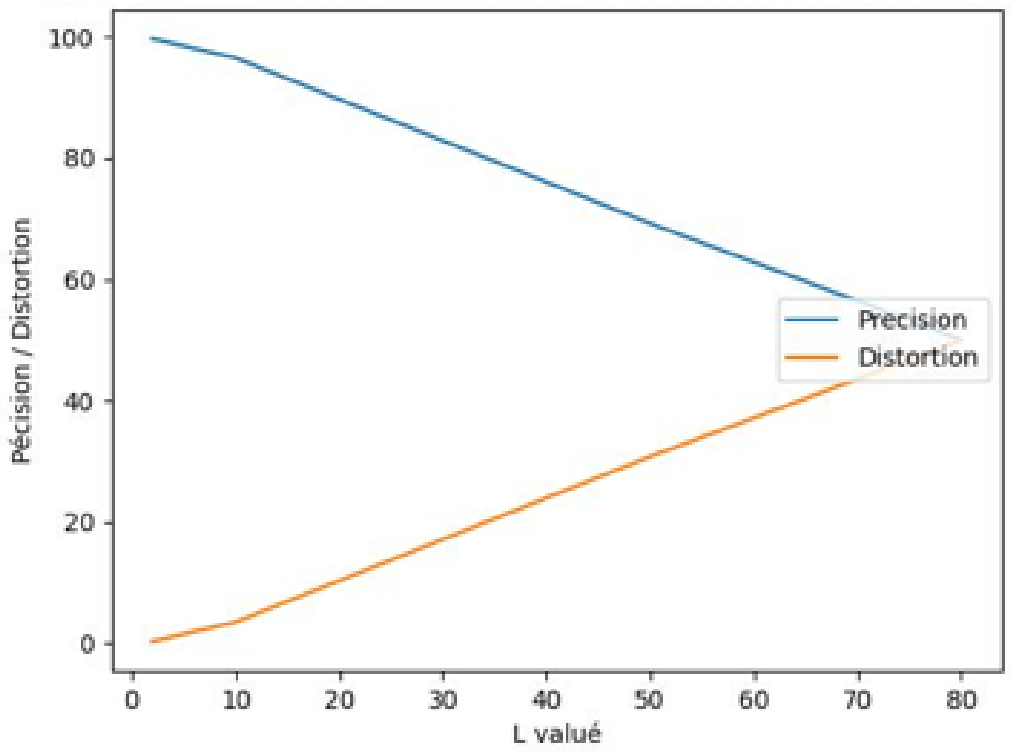

Figure 4.2: Diagramme de précision et distorsion en indemnités de l.

Observation :

— **Pour different l :** Cette section présente les résultats de l'anonymisation avec l, ce qui signifie que chaque groupe d'attribut sensibles doit avoir des différents attributs par rapport aux l [2, 10, 50, 80]. On remarque que plus la valeur du déve l augmente (par exemple, l=80) on voie un manque de données dans la taille du bdd anonymiser.

— **Pour la précision :** on remarque que plus la valeur du l augmente la précision démunit par exemple " l = 2 la précision = 99.69, pour l =50 la précision= 69.23 "

— **Pour la distortion :** on observe c'est l'inverse de la précision plus l est élever plus la distorsion augmente.

— **Pour le taux d'échec :** on voie une augmentation dans le taux d'échec par rapport aux l.

4.4.2 La base de données « pima »

• **Pour algorithme k-anonymat :** On a appliqué l'anonymisation avec l'algorithme k-anonymat sur la bdd « pima », avec plusieurs seuils k = [5, 50, 100, 500].

K	[2-40]	[50-80]	[100-400]	[500-600]
Précision	0.67	0.67	0.71	0.41
Distorsion	0.33	0.33	0.33	0.67
Taux d'échec	0.13	0.13	11.59	11.59
taille de bdd	768	768	680	680

Table 4.6: Résultats de test d'algorithme « k-anonymat » avec différant seuil k du bdd" pima ".

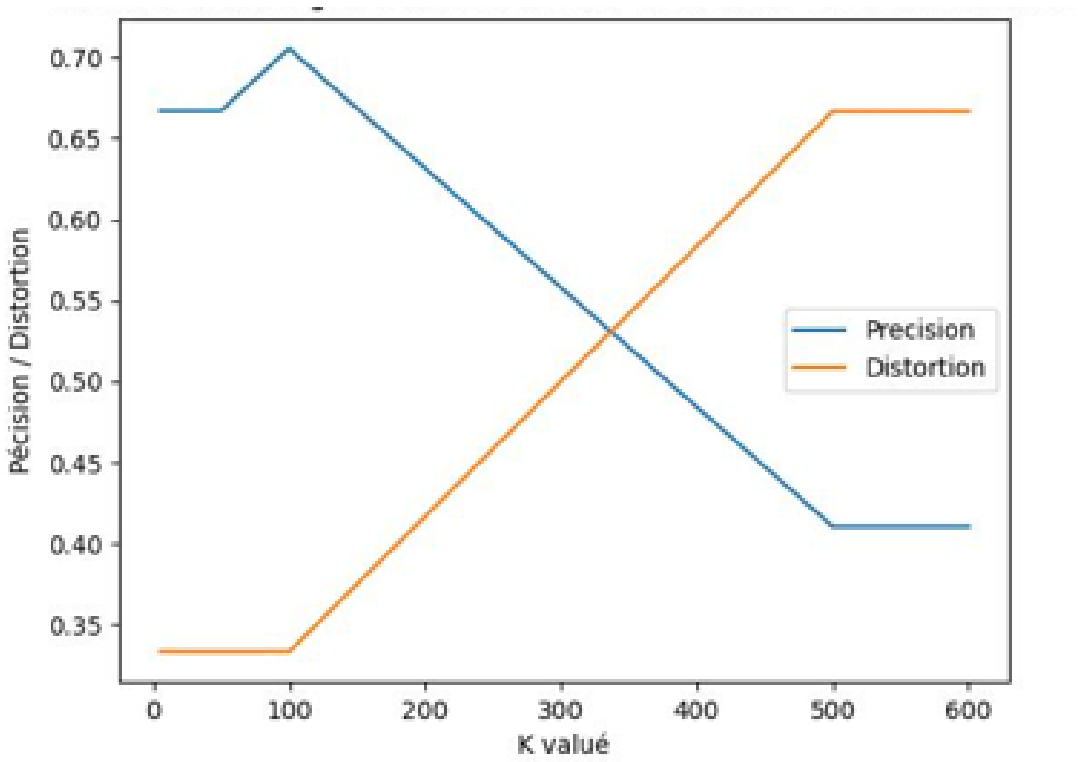

Figure 4.3: Diagramme de précision et distorsion en indemnités de seuil k.

Observation :

— **Pour plusieur seuil k :** plus le seuil k augmente, la taille de la base de données démunit, cela signifie une perte de données.

— **Pour la précision :** on voit une perturbation dans les valeurs de précision.

— **Pour la distorsion :** plus le seuil k augmente, il a une augmentation dans la distorsion.

— **Pour le taux d'échec :** on voit un taux d'échec élever par rapport au changement de seuil k.

Exemple : pour k=5.

Pregnancy	Glucose	BP	...	Insulin	MBI	Diabets	Age	Class
6	148	72	...	*	33.60	0.63	[50-75]	Y*
1	85	66	...	*	26.60	0.35	[25-50]	N*
8	183	64	...	*	23.30	0.67	[25-50]	Y*
1	89	66	...	*	28.10	0.17	[0-25]	N*
0	137	40	...	*	43.10	2.29	[25-50]	Y*
...	...	...	...	...	...	...	...	...

Table 4.7: bdd " pima " anonymisée par l'algorithme k-anonymat.

• **Pour l'algorithme l-diversity :** On appliqué l'anonymisation avec l'algorithme l-diversity sur la bdd « pima », avec plusieurs l.

L	[2-40]	[50-80]	[100-400]	[500-600]
Précision	46.42	0	0	0
Distorsion	53.58	100.0	100.0	100.0
Taux d'échec	53.58	100.0	100.0	100.0
Taille de bdd	357	0.0	0.0	0.0

Table 4.8: Résultats de test d'algorithme « l-diversity » avec différant l du bdd " pima ".

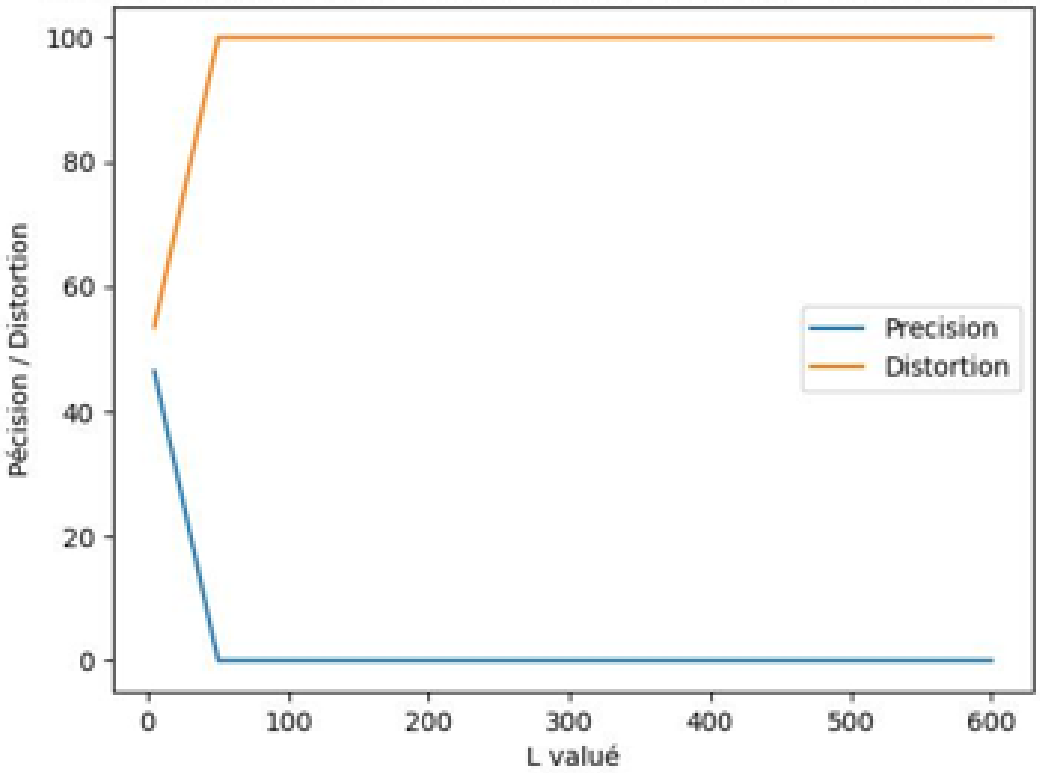

Figure 4.4: Diagramme de précision et distorsion en indemnités de l.

Observation :

— **Pour l = [5, 50, 100, 500] :** plus l augmente, la taille de la base de données démunit jusqu'à 0, cela signifie la perte de la base de données.

— **Pour la précision :** on voit la précision démunit par rapport à l'augmentation du l.

— **Pour la distorsion :** plus l augmente, il a une augmentation dans la distorsion.

— **Pour le taux d'échec :** on voit un taux d'échec élever par rapport au changement de l.

4.5 Discussions :

— Alors pour l'algorithme k-anonymat, on déduit que les mesures de précision indiquent que plus la valeur de seuil k est élevée, plus la précision diminue, cela ne signifie pas un bon anonymat, aussi une précision élevé signifie que les données sont anonymisées de manière plus précise et plus complète, cela peut être important pour protéger la vie privée des utilisateurs et éviter la divulgation d'informations sensibles.

— Il y a une grande distorsion des données par rapport au changement de seuil k, une distorsion peut aider à préserver la confidentialité peut être fait en ajoutant du bruit ou modifiant les valeurs des données de manière contrôlée pour réduire la corrélation entre les données et les individus. Une distorsion élevée ne signifie pas nécessairement un bon anonymat. En fait, une distorsion élevée peut indiquer que les données anonymisées sont très différentes des données originales, ce qui peut potentiellement compromettre la qualité et l'utilité des données anonymisées.

— Tel que le changement de seuil k est proportionnel à l'augmentation du taux d'échec, cela ne signifie pas un bon anonymat. Dans l'algorithme « k-anonymat » le seuil k élevé offre une meilleure protection des données, mais il peut y avoir une perte d'informations précises dans les données anonymisées.

— Affirme que l'algorithme l-diversity obtient un taux d'échec et une précision plus élevés après ce dernier démunit jusqu'à 0, ce qui augmente la distorsion.

— Donc il est important de trouver un équilibre entre l'anonymat des données et la précision et la distorsion nécessaire pour l'analyse ou l'utilisation des données, et avoir un mois taux d'échec.

— On déduit aussi que deux algorithmes (k-anonymat et l-diversité) garantissent un bon anonymat pour les bases de données grandes, car les petites bases présentent une perte de données excessive et un taux d'échec élevé.

• **Pour l'algorithme ϵ-Differential Privacy :** On a appliqué l'anonymisation avec l'algorithme ϵ-Differential Privacy sur la bdd " adult". La confidentialité différentielle est une définition mathématiquement rigoureuse de la confidentialité adaptée à l'analyse de grands ensembles de données. A alide de mécanique de laplace. Nous avons défini la donnée sensible capital-gain qui représente pour chaque individu et age, eduction, occupation, relationship comme des quasi-identifiers. On obtient une base de données nommée " anonymized_adult_data " :

	Age	Eduction	occupation	relationship	Class
0	13.308	5.0575	8.217	1.744	<=50k
1	24.497	11.977	1.876	2.421	<=50k
2	35.620	11.150	8.967	0.44	<=50k
3	14.1878	7.12	2.76	1.637	<=50k
4	17.064	10.813	8.61	1.924	<=50k
...	...	...	...	...	...

Table 4.9: bdd « adult » anonymisée par l'algorithme ϵ-Differential Privacy.

Observation :

— **Le taux de précision** = 74.42, Cela signifie que les données anonymisées conservent 72.42% des caractéristiques présentes dans les données originales.

— **SizeN :** 30162. Représente la taille de base de données non anonymisée.

— **SizeS :** 6033. Représente la taille de base de données après l'anonymisation.

— **On a calculé le taux d'échec (Hidden failure) :** En premier calculer m1 représente le nombre de ligne non masquée, et m2 représente le nombre de ligne masqué alors :
m1= 120 , **m2**=80.
L'échec =120, cela signifie 60% d'échec cachées.

On déduit que la taille de la base de données « adult.csv » a diminué. En général, une précision élevée est souhaitable car cela signifie que les données anonymisées conservent une grande similarité avec les données originales. Une précision élevée indique que les caractéristiques et les structures des données sont préservées dans une large mesure. Par contre le taux d'échec caché est élevé qui signifie généralement qu'une proportion importante des données anonymisées ne respecte pas les critères d'anonymisation ou présente des vulnérabilités potentielles pour la réidentification des individus.

4.6 Présentation de notre d'application :

Dans cette partie, nous présenterons les étapes de notre application.
Interface d'acceuil d'application permet d'accèder a toutes ses principales fonctionnalités. Elle contient les boutons :

— **Load data :** le bouton pour le chargement de la base de données.

— **Netoyage data :** le bouton pour faire le nettoyage de la base de données.

— **Insert k :** le bouton pour insert le nombre de seuil k.

— **Insert l :** le bouton pour insert le nombre de l.

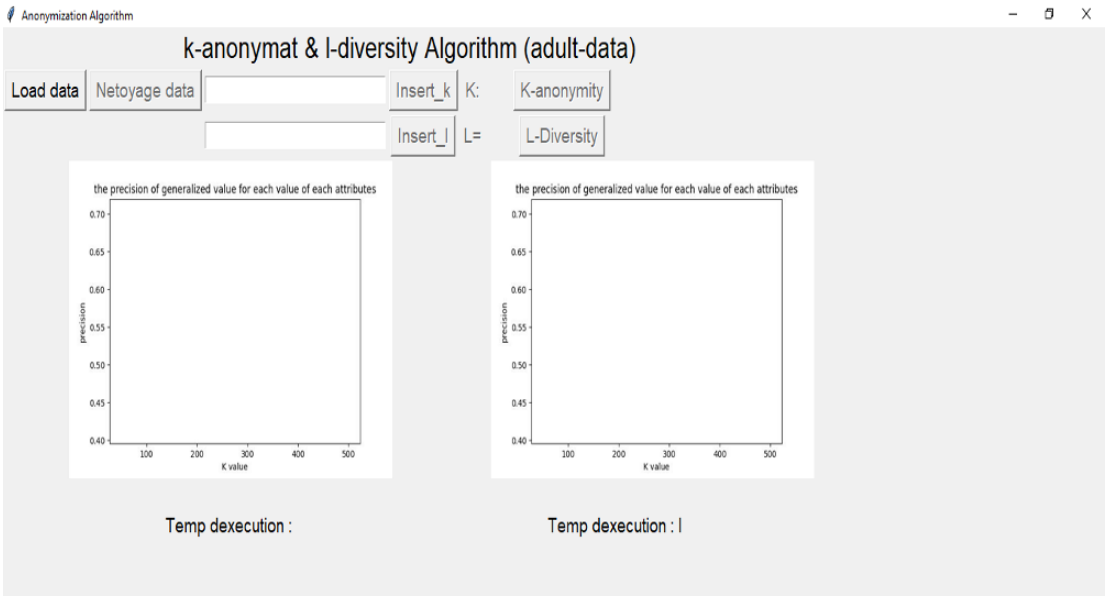

Figure 4.5: l'interface de l'application.

— Chargement de la base de données

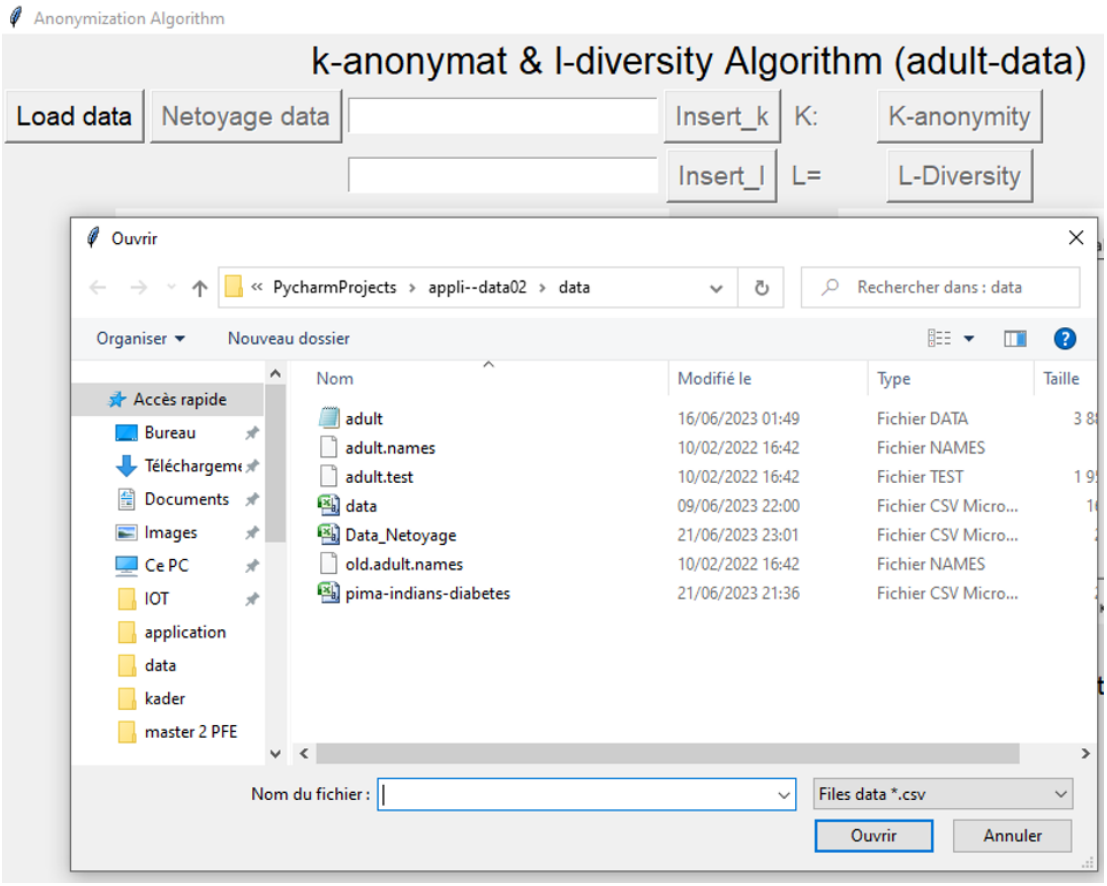

Figure 4.6: charger la base de données.

— Faire un nettoyage de la base de données, et insert le seuil k.

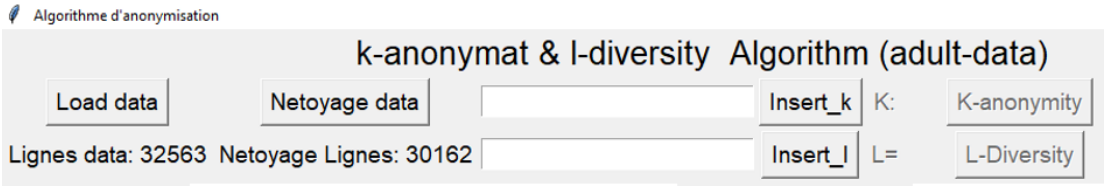

Figure 4.7: les boutons pour faire nettoyage et insert le seuil k et l.

— Affichege de résultat

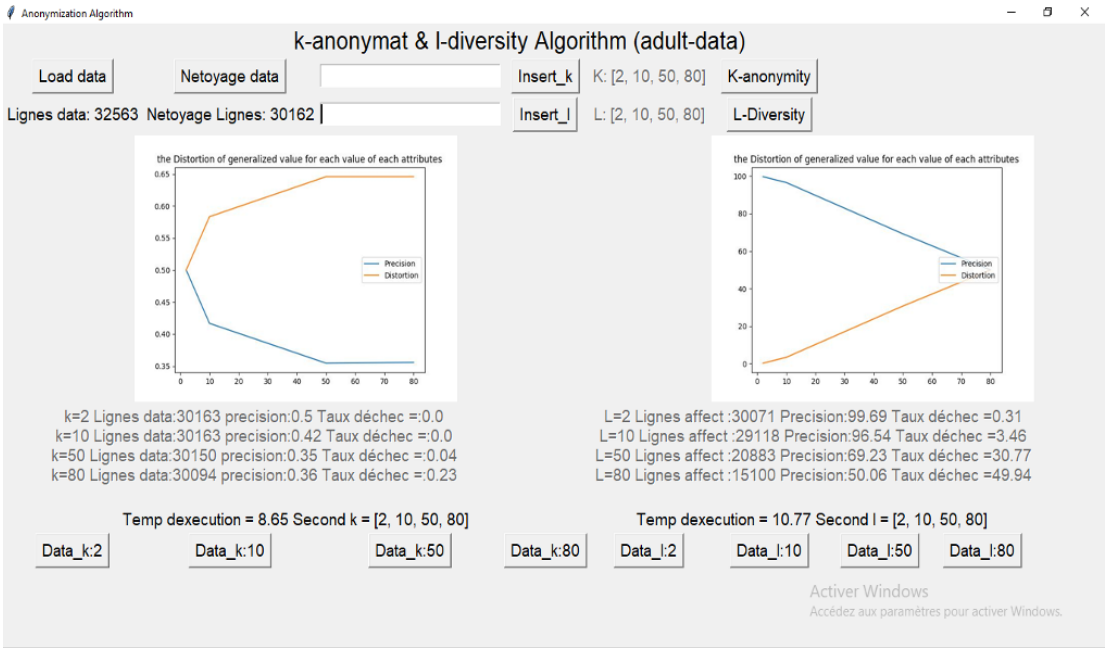

Figure 4.8: Résultats finale de l'application.

— **Data-k :2 :** le bouton pour afficher la base de données anonymiser avec l'algorithme k-anonymat tel-que k=2.

— **Data-l :2 :**le bouton pour afficher la base de données anonymiser avec l'algorithme l-diversity tel-que l=2.

4.7 Conclusion

l'anonymisation des données est une technique qui consiste à supprimer ou à modifier les informations personnelles d'un utilisateur afin de les rendre anonymes. Cela permet de protéger la vie privée de l'utilisateur et de réduire les risques de violation de données.
Alors dans notre travail on conclut que, la précision d'algorithme ϵ-Differential Privacy plus grande par rapport a l'algorithme k-anonymat et l-diversity, ce qu'il indique l'avantage d'informations sensibles sont préservées dans les données anonymisées, ce qui est également souhaitable. Alors que dans l'algorithme k-anonymat à la mesure d'une distorsion plus faible indique les données d'origine sont moins altérées lors de l'anonymisation.

Par contre un taux d'échec élevé dans l'algorithme ϵ-Differential Privacy mais beaucoup plus élever dans k-anonymat et l-diversity, qui peut résulter de divers facteurs, tels que des techniques d'anonymisation inefficaces, une mauvaise gestion des données sensibles ou des vulnérabilités potentielles dans le processus d'anonymisation. Cela peut entraîner des risques accrus pour la confidentialité des individus et une possibilité de réidentification.

En résumé, il est important de considérer que l'anonymisation des données est souvent réalisée pour protéger la vie privée des individus, donc il est important de trouver un équilibre entre la précision et la distorsion et avoir un taux d'échec faible.

Conclusion générale et perspectives

L'anonymisation des données sensibles est une pratique courante qui implique la suppression ou la modification d'informations. Cette pratique est essentielle pour protéger la vie privée des individus et prévenir la divulgation non autorisée de données personnelles. Étudier l'anonymisation des données sensibles dans les data sets revêt une grande importance, car cela permet de comprendre les meilleures pratiques pour protéger la vie privée des individus. Cela permet également aux organisations de développer des politiques et des processus efficaces pour la protection des données sensibles.

Cependant, l'anonymisation des données sensibles dans une base de données est un processus complexe qui nécessite des compétences techniques approprié pour être effectué correctement. Il existe de nombreux algorithmes et techniques d'anonymisation, mais choisir le bon algorithme peut être un défi qui requiert une expertise technique. Il est important de comprendre que l'anonymisation des données peut être délicate ainsi qu'elle exige une expertise technique pour être réalisée correctement.

Dans le cadre de notre projet de fin d'études, nous avons téster les étapes et l'application des techniques d'anonymisations, qui répond aux exigences de la protection des données, afin de déterminer le plus efficace pour le processus d'anonymisation. Notre mémoire est structuré de manière à couvrir l'importance de protéger la vie privée des personnes, les outils de publication de données, une présentation détaillée des algorithmes d'anonymisation utilisés dans notre projet, ainsi que la présentation des résultats obtenus.

En résumé, la protection de la vie privée et l'anonymisation des données sensibles dans les ensembles de données sont des enjeux cruciaux à l'ère de l'explosion des données. La mise en œuvre de bonnes pratiques d'anonymisation est importante pour prévenir les atteintes à la vie privée et garantir la confidentialité des informations personnelles. Il est essentiel de poursuivre la recherche et le développement d'algorithmes et de techniques d'anonymisation car il est extrêmement

difficile de nos jours d'obtenir une anonymisation parfaite pour les données sensibles. Même avec des méthodes d'anonymisation avancées, il reste toujours un risque de réidentification des personnes.
L'amélioration de cette application et trouver un bon équilibre entre la précision et la distortion fera l'objectif de nos études et recherches postérieur.

Bibliographie

[1] Calculer k-anonymat pour un ensemble de données | Documentation Cloud DLP | Google Cloud — cloud.google.com. https://cloud.google.com/dlp/docs/compute-k-anonymity?hl=fr. [Accessed 07-Jun-2023].

[2] Centre National de Ressources Textuelles et Lexicales — cnrtl.fr. https://www.cnrtl.fr/definition/sÃl'curitÃl'. [Accessed 12-april-2023].

[3] Confidentialité ; définition — marche-public.fr. https://www.marche-public.fr/Terminologie/Entrees/Confidentialite.htm. [Accessed 1-Jun-2023].

[4] Définition de attribut | Dictionnaire français — lalanguefrancaise.com. https://www.lalanguefrancaise.com/dictionnaire/definition/attribut. [Accessed 17-Jun-2023].

[5] Définition de tuple | Dictionnaire français — lalanguefrancaise.com. https://www.lalanguefrancaise.com/dictionnaire/definition/tuple. [Accessed 17-Jun-2023].

[6] POLITIQUE DE CONFIDENTIALITE - DEF Algerie — defalgerie.com. https://www.defalgerie.com/politique-de-confidentialite/. [Accessed 04-mars-2023].

[7] Pourquoi et comment pseudonymiser dans lapos ;administration | guides.etalab.gouv.fr — guides.etalab.gouv.fr. https://guides.etalab.gouv.fr/pseudonymisation/pourquoi-comment/#qu-est-ce-que-la-pseudonymisation. [Accessed 04-april-2023].

[8] Python (langage) — Wikipédia — fr.wikipedia.org. https://fr.wikipedia.org/wiki/Python_(langage). [Accessed 29-mai-2023].

[9] Python : définition et utilisation de ce langage informatique — journaldunet.fr. https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/. [Accessed 07-Jun-2023].

[10] Quelles techniques d'anonymisation pour protéger vos données personnelles ? - Octopize - Mimethik Data — octopize-md.com.

https://octopize-md.com/fr/2021/07/28/quelles-techniques-d'anonymisation-pour-proteger-vos-donnees-. [Accessed 04-march-2023].

[11] Anas Abou El Kalam, Yves Deswarte, Gilles Trouessin, and Emmanuel Cordonnier. Une démarche méthodologique pour l'anonymisation de données personnelles sensibles. In *2ème Conférence Francophone en Gestion et Ingénierie des Systèmes Hospitaliers*, 2004.

[12] Lydia Belhoul, Feriel Lalaoui, Malika Yaici, et al. *Anonymat et vie privée dans la blockchain.* PhD thesis, Université Abderrahmane Mira-Bejaia, 2021.

[13] Sabrina Khoulalène Bhavya Aggarwal. 7 méthodes de chiffrement des données pour garder les informations sensibles à l'abri des regards indiscrets. https://www.getapp.fr/blog/3552/methodes-chiffrement-donnees. [Accessed 20-mai-2023].

[14] Sylvain Castagnos. *Modélisation de comportements et apprentissage stochastique non supervisé de stratégies d'interactions sociales au sein de systèmes temps réel de recherche et d'accès à l'information.* PhD thesis, Université Nancy II, 2008.

[15] Heather Devane. Everything You Need to Know About K-Anonymity — immuta.com. https://www.immuta.com/blog/k-anonymity-everything-you-need-to-know-2021-guide/. [Accessed 04-Jun-2023].

[16] Pr Omar EL BAQQALI, Pr Abdelfettah SEDQUI, Pr Younès BENNANI, France USPN, Pr Olivier BODINI, Guénaël CABANES, and Pr Abdelouahid LYHYAOUI. Data anonymisation through unsupervised learning.

[17] Feten Ben Fredj. *Méthode et outil d'anonymisation des données sensibles.* PhD thesis, Conservatoire national des arts et metiers-CNAM ; Université de Sfax (Tunisie ..., 2017.

[18] Allannah Furlong. Cadre et confidentialité. *Filigrane*, 14(2) :62–76, 2005.

[19] Futura. Définition | Confidentialité | Futura Tech — futura-sciences.com. https://www.futura-sciences.com/tech/definitions/tech-confidentialite-1702/. [Accessed 10-Jun-2023].

[20] Omar Haboussi and Kheireddine Guenoune. *Protection de la vie privée (anonymat) dans les réseaux sociaux.* PhD thesis, université akli mohande-oulhadj bouira, 2019.

[21] Sahnine Mohamed. *Un anonymiseur web pour l'anonymisation des données personnelles sensibles.* PhD thesis, Université Oran1 Ahmed Ben Bella , Oran, 2022.

[22] Benjamin Nguyen. Techniques d'anonymisation. *Statistique et société*, 2(4) :53–60, 2014.

[23] DPO Partagé. T-Mobile : 37 millions de données sensibles volées — dpo-partage.fr. https://www.dpo-partage.fr/t-mobile-violation-donne{é}s/. [Accessed 20-Jun-2023].

[24] Vincent Thouvenot, Thibaut Dubois, and Stephane Lorin. Anonymisation et confidentialité différentielle appliquéesa des données spatio-temporelles : cas d'usage portant sur la billettique.