


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Laboratoire de recherches LIMPAF

Mémoire de Master

en Informatique

Spécialité : *Ingénierie des Systèmes d'Information et Logiciel*

Thème

Privacy-preserving classification

Encadré par

— Mr. Boudjelaba Hakim

Réalisé par

— Mlle. messaour Imane

— Mlle. guellil Yasmine

Les jurys

— Mme. AID Aicha	Présidente
— Mr. OUKAS Noureddine	Examinateur1
— Mme. SAOUD Afaf	Examinateur2

2021/2022

Remerciements

Nous tenons tout d'abord à remercier ALLAH le tout puissant, qui nous a donné la force, la capacité et surtout la patience pour accomplir ce travail.

Nous tenons à adresser nos plus chaleureux remerciements à Monsieur « BOUDJELABA Hakim », notre promoteur de ce PFC, pour son aide illimitée et son soutien pendant toute cette période pleine de défis. Ses conseils avisés et sa patience nous ont aidés à surmonter l'hésitation, l'embarras et de diriger le projet et avoir des meilleurs résultats. Sa constante disponibilité à notre égard et de nous avoir donné l'occasion de travailler sur un tel sujet de recherche qui est très riche d'information et de découverte pour nous.

Nous tenons à exprimer nos sincères remerciements à tous les professeurs de l'Université Akli Mohand Oulhadj qui m'ont enseigné et qui par leurs compétences et sérieux nous ont permis de poursuivre nos études.

Nous tenons également à remercie notre chère amie Amira de ses précieuse aide qu'il nous a apportée et pour tous ses conseils utiles durant la période de notre projet.

Nos vifs remerciements iront aussi aux membres de jury qui nous ferons l'honneur de juger notre travail et de soulever les critiques nécessaires affin d'enrichir nos connaissances et d'apporter un plus à notre travail.

Nos remerciements s'adressent aussi à toutes les personnes qui nous ont soutenu, de prêt ou de loin, durant la réalisation de cette PFC.

Merci à tous.

Dédicaces

À ma chère mère, vous avez su porter en moi les soins et consentir les efforts pour mes études. Aucune dédicace ne saurait exprimer tout le respect et l'amour que je vous porte, vous m'avez toujours fait confiance. Veuillez trouvez en ce travail la consolation et le témoin de la patience.

À mon cher père, malgré les grandes responsabilités que vous assummez, vous avez toujours été près de moi, pour me soutenir, me suivre, et m'encourager. Puisse ce travail diminuer votre souffrance et vous porter bonheur.

À mes chers frères et sœurs, je vous remercie pour tous les bienfaits que chacun a pu faire pour moi.

À ma chère binôme et meilleure amie, merci pour votre patience et votre soutien dans toutes les difficultés que nous avons rencontrées, que Dieu vous accorde tout ce que vous souhaitez.

À toute la famille et tous ceux qui m'aiment et que j'aime.

Messaour Imane.

Dédicaces

Je ne peux pas commencer sans mentionner le nom d'ALLAH que je remercie pour tout et avant tout, et grâce à lui j'ai pu réaliser ce travail, que j'ai didié :

À la lumière qui a illuminé mon chemin et qui a fait des efforts au fil des ans pour gravir les échelons du succès (mon cher père) et à qui Dieu a donné le ciel sous ses pieds et m'a comblé d'amour et de tendresse et m'a fait me sentir heureuse et en sécurité ma mère (Djamila) sans eux je n'aurais pas lu un livre ou appris une lettre.

À ceux qui ont été ma forteresse, mon cœur fermé et la saveur de la vie, à ceux qui attendent jour après jour ma graduation, mes frères et sœurs Hakim Brahim Khaled Wahiba et Khadidja.

À la personne la plus précieuse que je suis fière de connaitre Walid.

Pour qui était un titre de soutien et d'amour, nous avons traversé les moments les plus difficiles ensemble, ma chère binôme, ou je dis ma soeur imane, merci pour ce chemin qui m'a rapproché de toi.

À tous ceux qui m'ont soutenu de près ou de loin, merci.

Guellil Yasmine.

Résumé

La classification préservant la vie privée (développer des modèles sans voir les données) est importante pour l'apprentissage automatique et l'exploration de données, les sources de données collaborent pour développer un modèle global, mais ne doivent pas divulguer leurs données à d'autres. Dans de nombreuses situations, les données sont réparties entre plusieurs organisations. Ces organisations peuvent vouloir utiliser toutes les données pour créer des modèles prédictifs plus précis tout en ne révélant ni leurs données/bases de données d'apprentissage ni les instances à classer.

A travers ce mémoire, nous avons proposé une solution originale et sécurisée, incluant le respect de la vie privée. Nous avons anonymisé notre jeu de données en ajoutant des données fectives pour perturber l'ensemble de données original tout en essayant de ne pas perdre l'utilité des données originales, puis nous avons évalué notre proposition avec l'ensemble de données orginal et on a comparé les résultats de classification entre le jeu de données original et perturbé.

Mots clés : Classification, Vie privée, Anonymat, Gaussian naive Bayes, Apprentissage automatique, Perturbation.

Abstract

Privacy-preserving classification (developing models without seeing the data) is important for machine learning and data mining, data sources collaborate to develop an overall model, but should not disclose their data to others others. In many situations, data is distributed across multiple organizations. These organizations may want to use all the data to build more accurate predictive models while not revealing their training data/databases or instances to classify.

Through this memory, we have proposed an original and secure solution, including respect for privacy. We anonymized our dataset by adding fictitious data to perturb the original dataset while trying not to lose the utility of the original data, then we evaluated

our proposal with the orginal dataset and we got compared the classification results between the original and perturbed dataset.

ملخص

يعد تصنيف الحفاظ على الخصوصية (تطوير النماذج دون رؤية البيانات) أمرًا مهمًا للتعلم الآلي واستخراج البيانات ، حيث تتعاون مصادر البيانات لتطوير نموذج شامل ، ولكن لا ينبغي الكشف عن بياناتها للآخرين. في كثير من الحالات ، يتم توزيع البيانات عبر مؤسسات متعددة. قد ترغب هذه المؤسسات في استخدام جميع البيانات لبناء نماذج تنبؤية أكثر دقة مع عدم الكشف عن بيانات التدريب \ قواعد البيانات أو الحالات التي يجب تصنيفها.

من خلال هذه المذكرة ، اقترحنا حلاً أصليا وآمنا يحترم الخصوصية ، حيث قمنا بإخفاء هوية مجموعة البيانات الخاصة بنا عن طريق إضافة بيانات وهمية لتشويش مجموعة البيانات الأصلية دون فقدان فائدة البيانات ، ثم قيمنا اقتراحنا بمقارنة نتائج التصنيف بين مجموعة البيانات الأصلية والمشوشة.

الكلمات المفتاحية : التصنيف ، الخصوصية ، عدم الكشف عن الهوية، التعلم الآلي ، التشويش

Table des matières

Table des matières **1**

Table des figures **4**

Liste des tableaux **7**

Liste des abréviations **8**

Introduction générale **1**

1 Vie privée et anonymat **3**

1.1 Introduction 3

1.2 La vie privée 3

1.3 L'anonymat 4

1.4 Différence entre *vie privée* et *anonymat* 4

1.5 Relation entre *vie privée* et *anonymat* 5

1.6 La problématique d'anonymat 5

1.7 Les techniques d'anonymat 5

1.7.1 La généralisation 5

1.7.2 La randomisation 6

1.7.3 La substitution 6

1.7.4 La permutation 6

1.7.5 La perturbation 7

1.8 Les modèles de protection de la vie privée 7

1.8.1 La pseudonymisation 7

1.8.2 k-anonymat 8
1.8.3 l-diversité 11
1.8.4 t-proximité 12
1.8.5 Confidentialité différentielle 13
1.9 Conclusion 14

2 Classification et anonymat **15**
2.1 Introduction 15
2.2 La classification 15
2.3 Méthodes de classification 16
2.3.1 L'arbre de décision 16
2.3.2 Classification naïve bayésienne 18
2.3.3 Support vecteur machine (SVM) 20
2.3.4 K nearest neighbors 23
2.3.5 Les réseaux de neurones artificiels 24
2.4 Anonymat pour la classification 26
2.4.1 La distribution des données 26
2.4.2 Les approches de classification préservant l'anonymat 28
2.4.3 Étude comparative des techniques precédente 38
2.5 Conclusion 40

3 Approche proposée et validation **41**
3.1 Introduction 41
3.2 Problématique 41
3.3 Architecture de notre approche 42
3.4 Nettoyage et pré-traitement des données 43
3.4.1 Création du jeu de données(La matrice de données) 43
3.4.2 La description des variables de fichier "heart" 44
3.4.3 L'exploration et visualisation des données de heart 44
3.4.4 Nettoyage et pré-traitement du jeu de données heart 46
3.5 Approche proposée 46
3.6 Explication de l'approche 47
3.7 Environnements et outils de développement 51

3.7.1 Plate-formes et environnement 51
3.7.2 Langage de programmation 52
3.7.3 Bibliothèques Utilisées 52
3.8 Resultats expérimentaux et évaluation 53
3.8.1 Matrice de confusion 54
3.8.2 Accuracy 56
3.8.3 Précision 57
3.8.4 Rappel(Recall) 58
3.8.5 F1 score 59
3.8.6 Entropie 60
3.8.7 Courbe ROC 61
3.8.8 Autres évaluation 62
3.9 Conclusion 64

Conclusion générale et perspectives **65**

Bibliographie **67**

Table des figures

1.1 Pseudonymisation et exemple de calcul [1] 8
1.2 Jeux de données anonymisées [2]. 9
1.3 Arbre de généralisation [2]. 10
1.4 Anonymisation en cours [2]. 10
1.5 Données anonymes [2]. 11
1.6 Tableaux données brutes & données anonymes et diverses [2]. 11

2.1 L'arbre généré par ID3 de l'exemple jouer Tennis. 18
2.2 Illustration de l'hyperplan séparateur et la marge maximale [3]. 21
2.3 Data set avec support vector machine [4]. 23
2.4 Prédiction d'un critère avec réseau de neurones [5]. 25
2.5 Exemple d'un système de données centralisées [6]. 26
2.6 Exemple d'un système des données distribuées [7]. 27
2.7 Distribution horizontale des données [8]. 27
2.8 Distribution verticale des données [8]. 28
2.9 Les approches de classification préservant l'anonymat. 29
2.10 Illustration en appliquant l'approche ID3 classique avec les échantillons originaux T_S [9]. 31
2.11 llustration en appliquant l'approche ID3 modifiée avec les échantillons irréalisés ($T^{'} + T^{P}$). Pour chaque étape, les valeurs d'entropie et les sous-arbres résultants sont exactement les mêmes que ceux de l'approche traditionnelle [9]. 32

3.1 Objectif de notre travail. 42

3.2 Architecture générale de notre proposition. 42
3.3 Dataset avant prétraitement. 43
3.4 Le rapport HTML du jeu de données heart. 45
3.5 Dataset après la supression des lignes dupliquées. 46
3.6 Architecture de notre proposition. 47
3.7 Partie du dataset original qui a juste la classe 0. 48
3.8 Partie du dataset original qui a juste la classe 1. 48
3.9 Le maximum et le minimum des valeur pour chaque colonne dans les deux jeux de données. 49
3.10 Les lignes ajoutées. 49
3.11 Dataset anonymisée. 50
3.12 La division de l'ensemble de données verticalement et horizontalement. . . 51
3.13 Matrice de confusion du jeu de données original. 55
3.14 Matrice de confusion du jeu de données anonymisé. 55
3.15 Secteur des proportions des valeurs de matrice de confusion (original). . . . 55
3.16 Secteur des proportions des valeurs de matrice de confusion (anonymisé). . 55
3.17 Les résultats des attributs d'évaluation du dataset original. 56
3.18 Les résultats des attributs d'évaluation du dataset anonymisé. 56
3.19 Comparaison d'accuracy entre le dataset original et anonymisé. 57
3.20 Comparaison de précision entre le dataset original et anonymisé. 58
3.21 Comparaison de rappel entre le dataset original et anonymisé. 59
3.22 Comparaison de F1 score entre le dataset original et anonymisé. 60
3.23 Comparaison d'entropie entre le dataset original et anonymisé. 61
3.24 Courbe ROC du dataset original. 62
3.25 Courbe ROC du dataset anonymisé. 62
3.26 Proportion des attributs de colonne cible (original). 62
3.27 Proportion des attributs de colonne cible (anonymisé). 62
3.28 Histogramme pour les probabilités prédites de 0 (original). 63
3.29 Histogramme pour les probabilités prédites de 0 (anonymisé). 63
3.30 Histogramme pour les probabilités prédites de 1 (original). 63
3.31 Histogramme pour les probabilités prédites de 1 (anonymisé). 63
3.32 Prédiction avec le dataset original. 64

3.33 Prédiction avec le dataset anonymisé. 64
3.34 Prédiction avec le dataset original. 64
3.35 Prédiction avec le dataset anonymisé. 64

Liste des tableaux

1.1 Jeux de données non anonymisées [2]. 9
1.2 t-proximité. 13

2.1 Ensemble d'exemple pour jouer le Tennis [10]. 17
2.2 Exemple d'une data set. 19
2.3 Exemple d'une Data set. 22
2.4 Exemple d'une Data set [11]. 24
2.5 Exemple d'une Data set. 25
2.6 Étude comparative des techniques precédentes. 39

Liste des abréviations

QI	Quasi-Identifiant
TI	Technologie de l'Information
IP	Internet Protocol
FAI	Fournisseur d'Accès à Internet
ID3	Itérative Dichotomiser 3
CART	Classification And Regression Trees
SVM	Support Vecteur Machine
RNA	Réseaux de Neurones Artificiels
SMC	Secure Multi-party Computation
BDD	Base De Données
TAN	Tree Augmented Naive Bayes
PP-SVM	Privacy-Preserving Multi-Class Support Vector Machine
PPSVC	Privacy-Preserving Support Vector Machine Classifier
GNB	Gaussian Naive Bayes
HTML	HyperText Markup Language
ROC	Receiver Operating Feature Curve

Introduction générale

Motivation et description de la problématique

Ces dernières années, les approches d'exploration de données ont été considérées comme un outil pour améliorer la productivité du marché des technologies de l'information (IT) en exploitant les connaissances extraites des données. Le partage d'informations dans une tâche d'exploration de données permet généralement d'obtenir des informations supplémentaires grâce à une analyse collaborative des informations, ces informations peuvent être exploitées pour améliorer les revenus, par exemple grâce à l'analyse du panier d'achat ou prévenir les pertes dues à de nouvelles cyber menaces potentielles. D'autres applications incluent l'analyse des données médicales fournies par de nombreux hôpitaux et centres de santé pour l'analyse statistique des dossiers des patients utiles pour la formation des causes et des symptômes liés à la nouvelle pathologie. Quel que soit l'objectif final, malheureusement, le partage d'informations s'accompagne de problèmes et d'inconvénients qu'il faut résoudre et les informations partagées peuvent être sensibles et potentiellement préjudiciables à la vie privée des individus, telles que les dossiers des employés pour les applications commerciales ou les dossiers des patients à des fins médicales. Le maintien de la confidentialité de nos jours est devenu un revers pour le stockage des données dans un entrepôt de données, c'est pourquoi les données dans l'entrepôt doivent être rendues indiscernables, et les données sont cryptées pendant le stockage et décryptées plus tard si nécessaire à des fins d'analyse dans l'exploration de données.

Objectif

L'objectif est de publier des données anonymes générées à partir de données brutes qui protègent contre les risques de ré-identification. En d'autres termes, il ne devrait pas être

possible de trouver dans les données publiées un individu qui se trouve dans les données originales.

Contributions

Implémentation de l'approche proposée en utilisant les techniques d'anonymat ainsi un algorithme de classification, ensuite l'évaluation des resultats obtenues de classification en comparant divers mesures d'évaluation entre le jeu de données obtenu et l'original.

Organisation du manuscrit

Pour bien illustrer nos contributions dans le cadre de ce travail, le rapport est organisé comme suit :

Le chapitre 1 est consacré au définitions des techniques d'anonymat et vie privé et les différentes méthodes utilisées pour protéger l'anonymat.

Le chapitre 2 est consacré aux définitions des méthodes de classification, ensuite une étude de quelques travaux connexes à notre domaine de recherche.

Le chapitre 3 présente en détail notre approche utilisée pour l'anonymat des données ainsi l'algorithme de classification choisi pour l'implémentation, et on termine le chapitre par les différents outils et environnement de développement utilisés dans notre étude et illustrer les résultats obtenues.

Et enfin, nous clôturons notre travail par une conclusion générale et quelques perspectives.

Chapitre 1

Vie privée et anonymat

1.1 Introduction

L'anonymat garantit la liberté d'expression et préserve la vie privée, elle protège les particuliers des représailles potentielles des gouvernements, des multinationales, des groupes de pression ou des organisations criminelles.

Dans ce chapitre, on va définir en premier lieu la vie privée et l'anonymat, puis présenter la différence et la relation entre ces deux dernières, par la suite nous allons présenter les différentes méthodes utilisées pour protéger la vie privée et mentionner les limites de chacune, et on va terminer cette partie en représentant les différentes techniques d'anonymat.

1.2 La vie privée

La vie privée sur internet, également appelée confidentialité en ligne, est essentiellement la confidentialité personnelle à laquelle vous avez droit lorsque vous consultez, stockez ou fournissez des informations vous concernant sur Internet. Les entreprises, les gouvernements et même les pirates s'intéressent à vos informations pour différents motifs, tels que les identifiants, les habitudes de navigation, les intérêts personnels ou les opinions, c'est pourquoi les données personnelles et la vie privée sont protégées à l'ère des réseaux sociaux comme Facebook et Twitter, des smartphones et des appareils connectés [12].

Les données personnelles sont l'essence de l'économie numérique, les sociétés numériques s'enrichissent grâce aux données récoltées sur les internautes, malheureusement on a pu

constater avec de nombreux scandales que les sociétés ne sont pas en mesure de garantir à 100% la sécurité des données de leurs utilisateurs. Les gouvernements et les entreprises dressent le profil des internautes, et si ces informations finissent entre de mauvaises mains, il serait théoriquement possible de les manipuler pour changer leurs façons de penser ou même de voter [13].

1.3 L'anonymat

L'anonymat global désigne l'état de quelqu'un ou de quelque chose qui choisit de rester anonyme, inaccessible, impossible à suivre, et elle a traversé et développé à travers les époques et est directement influencé par les avancées techniques des moyens de communication et du droit en général.

Être anonyme en ligne ne consiste pas seulement à se cacher derrière un pseudonyme, la tâche est beaucoup plus complexe, le véritable anonymat nécessite l'intraçabilité, et évidemment la nature des réseaux informatiques rend cela difficile. Mais l'anonymat sur Internet est un besoin nécessaire et une liberté pour de nombreuses personnes, que ce soit pour affirmer leur différence sans crainte de représailles, pour parler de problèmes personnels ou de santé ...[13].

1.4 Différence entre *vie privée* et *anonymat*

La vie privée et l'anonymat sont deux termes fondamentaux dans le phénomène de collecte et de gestion des données qui implique la participation du public. Les termes vie privée et anonymat sont utilisés de manière interchangeable, et c'est faux. Par conséquent, une distinction doit être faite entre ces deux mots, en particulier dans les études de recherche auxquelles participent des humains [14].

La différence principale entre la vie privée et l'anonymat est que dans la confidentialité, seul le chercheur connaît l'identité des participants alors que dans l'anonymat, même le chercheur ne connaît pas l'identité des participants [14].

1.5 Relation entre *vie privée* et *anonymat*

L'anonymat et la vie privée sont souvent liés, et pour cause le premier est un moyen de mémoriser le deuxième. La vie privée est la raison de l'utilisation de la technologie anonyme, elle ne peut être comprise qu'en contrôlant ce que vous laissez sur Internet. Cela implique de garder le contrôle sur les informations personnelles et de ne pas les laisser sortir du cadre dans lequel elles sont rendues publiques [15].

1.6 La problématique d'anonymat

Les principaux points que les responsables du traitement des données doivent prendre en considération lorsqu'ils choisissent de mettre en œuvre une technologie particulière [16] :

- **Individualisation** : correspond à la capacité d'isoler tout ou une partie des enregistrements identifiant un individu ;
- **Corrélation** : consiste à la possibilité de relier au moins deux enregistrements liés à la même personne concernée ou au même groupe de personnes concernées (Soit dans la même base de données soit dans deux bases de données différentes) ;
- **Inférence** : est la possibilité de déduire, avec un haut degré de probabilité, la valeur d'un trait des valeurs d'un autre ensemble de traits.

1.7 Les techniques d'anonymat

Ils existent différentes techniques d'anonymisation. Lors d'un processus d'anonymisation, on peut décider d'en utiliser une ou plusieurs techniques selon le volume de données, les besoins, ...

1.7.1 La généralisation

La généralisation consiste à exclure délibérément certaines données pour les rendre moins identifiables. Les données peuvent être modifiées en une série de plages ou une grande région avec des limites raisonnables. Par exemple, le numéro de maison à une adresse peut être supprimé, mais assurez-vous que le nom de la voie n'est pas supprimé. L'objectif est de supprimer certains identifiants tout en préservant l'exactitude

des données [17].

Cette opération remplace certaines valeurs par une valeur parente dans la taxonomie d'un attribut [18].

Par exemple, la valeur 25 d'un âge quasi-identifiant peut être généralisée avec un intervalle [25–30] ou < 30. Cette opération repose sur la taxonomie de chaque QI, L'objectif est de supprimer certains identifiants tout en préservant l'exactitude des données [19].

Quasi-identifiant

Les quasi-identifiants sont des éléments d'information qui ne sont pas en eux-mêmes des identifiants uniques, mais qui sont suffisamment bien corrélés avec une entité pour pouvoir être combinés avec d'autres quasi-identifiants pour créer un identifiant unique. Les quasi-identifiants peuvent ainsi, lorsqu'ils sont combinés, devenir des informations d'identification personnelle. Ce processus est appelé ré-identification [20].

1.7.2 La randomisation

La randomisation est le processus de modification des caractéristiques d'un ensemble de données afin qu'elles soient moins précises, tout en préservant la distribution globale. Par exemple, modifier les données sur la date de naissance des individus pour changer l'exactitude des informations contenues dans la base de données [16].

1.7.3 La substitution

Les valeurs de données sont remplacées par des valeurs alternatives fausses, mais réalistes. Par exemple, les vrais noms de clients sont remplacés par une sélection aléatoire de noms à partir d'un annuaire téléphonique [21].

1.7.4 La permutation

Il s'agit de mélanger les valeurs d'attributs dans un tableau de manière à ce qu'elles soient artificiellement liées aux différentes personnes concernées. Ainsi, la commutation modifie les valeurs dans l'ensemble de données dès que vous les basculez d'un enregistrement à un autre. Exemple : après l'anonymisation, l'âge du patient A a été remplacé par l'âge du patient C [19].

1.7.5 La perturbation

Dans cette opération, les valeurs des données d'origine sont remplacées par des valeurs générées de manière synthétique. De plus, les valeurs synthétiques sont générées de telle sorte que les informations statistiques ne diffèrent pas beaucoup dans les deux ensembles de données (c'est-à-dire les ensembles de données réels et générés synthétiquement) [19].

1.8 Les modèles de protection de la vie privée

1.8.1 La pseudonymisation

Les pseudonymes sont un processus dans lequel les identifiants personnels sont remplacés par des pseudonymes au fur et à mesure que les données perdent leur caractère nominal Métadonnées indirectes (alias, numéro dans la taxonomie, etc.). Tout en protégeant la confidentialité de la personne concernée, les fausses données sont toujours considérées comme des données personnelles car elles sont susceptibles d'être retirées dans certains cas. En raison de ce crénelage, il doit s'exécuter sur des algorithmes complexes qui rendent la ré-identification aussi difficile que possible. Le grand avantage de l'utilisation de pseudonymisation est l'absence de restrictions sur le traitement supplémentaire des données. Tant qu'on a affaire directement à des champs non identifiables, on peut effectuer exactement les mêmes calculs qu'avec une base de données non identifiable [22].

La figure 1.1 représente la pseudonymisation et exemple de calcul.

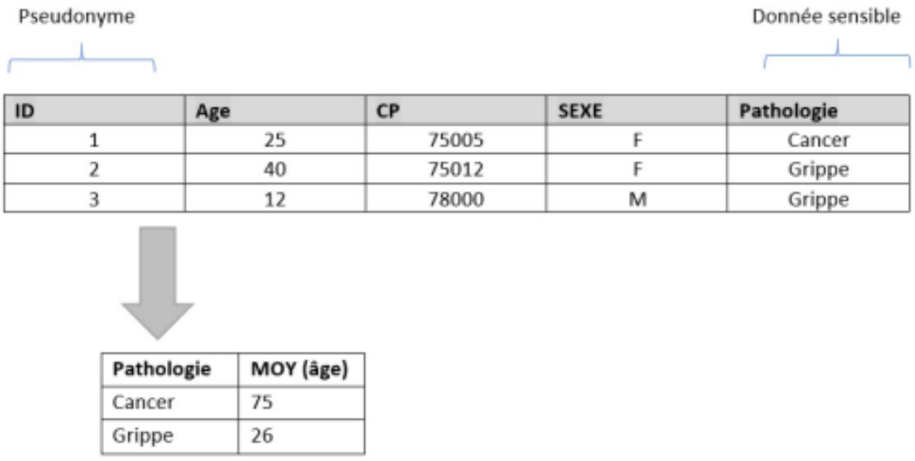

FIGURE 1.1 – Pseudonymisation et exemple de calcul [1]

Inconvénients

— La principale limite de l'utilisation de pseudonymes est qu'il reste techniquement possible de réattribuer des identifiants avec des pseudonymes à des personnes physiques ; Bien que cette opération soit généralement interdite et que des sanctions sévères lui soient infligées.

— L'utilité des données se détériore.

1.8.2 k-anonymat

La k-anonymisation est un concept clé introduit pour faire face au risque de réidentification des données anonymisées par association avec d'autres ensembles de données. Le modèle du k-anonymat a été proposé pour la première fois en 1998 par Latanya Sweeney dans son article "Privacy Protection in Information Disclosure : k-anonymity and Its Application through Generalization and Repression". Pour atteindre l'anonymat pour k, il doit y avoir au moins k individus dans l'ensemble de données qui partagent l'ensemble de traits qui peuvent devenir des identifiants pour chaque individu. L'anonymat k peut être décrit comme une garantie de " dissimulation dans la foule" : si tout le monde fait partie d'un groupe plus large, alors n'importe lequel des enregistrements de ce groupe peut correspondre à n'importe qui [22].

Si $k = 3$ alors nous pouvons faire deux groupes avec n nombre de groupes aléatoires (voir Table 1.1) : un groupe en bleu et un groupe en jaune.

Nom	**CP**	**Age**	**Poids**
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	75

TABLE 1.1 – Jeux de données non anonymisées [2].

Nous séparons ensuite ces informations en deux tableaux distincts : le tableau "quasi-identifiants" à gauche et le tableau "données sensibles" à droite (voir Figure 1.2).

CP	Age	Groupe
18000	22	*G1*
18500	21	*G1*
69000	27	*G1*
18510	20	*G2*
69100	26	*G2*
69300	28	*G2*

Table QID

Groupe	Poids
G1	50, 70, 90
G2	60, 70, 75

Table DS

FIGURE 1.2 – Jeux de données anonymisées [2].

Dans le tableau de gauche, les trois premières lignes correspondent au groupe 1 et les trois lignes suivantes correspondent au groupe 2.
Trois valeurs de poids sensibles sont associées à chacun de ces groupes [2] :

— Pour le G1, les trois valeurs sont respectivement 50, 70 et 90 kg.

— Pour le G2, les trois valeurs sont respectivement 60, 70 et 75 kg.

Ainsi, même si l'on connaissait le code postal et l'âge d'une personne, il est impossible de connaître le poids d'une personne, avec une certitude supérieure au tiers [2].

Inconvénients

— L'utilité des données peut être ambiguë, afin de résoudre ce problème d'ambiguïté, nous allons devoir aborder la perception de la donnée k anonymes par généralisation.

***k*-anonymat par généralisation**

Cette technique consiste à créer une hiérarchie de généralisation permettant de convertir des données brutes en données moins précises [2].

Faisons d'abord un arbre de généralisation pour chaque attribut (voir Figure 1.3). Si l'on prend le cas du code postal, on peut généraliser en 3 niveaux : un niveau département, un niveau région et un niveau national [2].

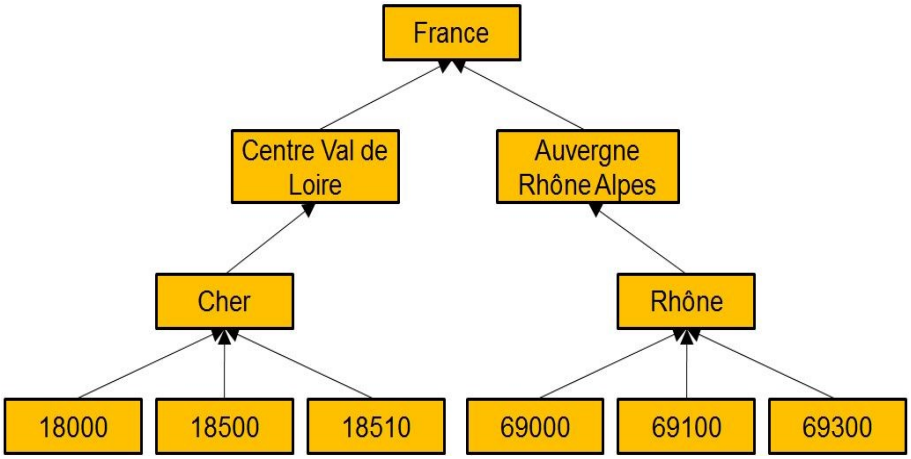

FIGURE 1.3 – Arbre de généralisation [2].

Une fois cet arbre de généralisation créé, nous allons généraliser la valeur de ces attributs jusqu'à ce que toutes les combinaisons correspondent à au moins $k-1$ d'autres attributs. Revenons au tableau des données brutes : si on généralise le code postal au niveau du département, les six groupes restent différents [2] (voir Figure 1.5).

CP	Age	Poids
Cher	22	50
Rhône	27	70
Cher	21	90
Cher	20	60
Rhône	26	70
Rhône	28	75

FIGURE 1.4 – Anonymisation en cours [2].

Alors, ce n'est pas encore fini ! Continuons à généraliser l'âge. Comme vous pouvez le constater, nous vous présentons deux périodes différentes de [20-24] ou [25-29] [2](Figure 1.5).

CP	Age	Poids
Cher	[20-24]	50
Rhône	[25-29]	70
Cher	[20-24]	90
Cher	[20-24]	60
Rhône	[25-29]	70
Rhône	[25-29]	75

FIGURE 1.5 – Données anonymes [2].

1.8.3 *l*-diversité

L'objet de la l-diversité est de s'assurer que chaque ensemble inconnu k est également suffisamment diversifié ou suffisamment varié, c'est-à-dire avec suffisamment de valeurs plausibles distinctes en son sein, afin d'empêcher l'inférence par homogénéité de la valeur plausible d'une personne. Il faut donc associer à chaque classe de k individus au moins l valeurs sensibles dites bien représentées. Cela signifie qu'il y a suffisamment de valeurs différentes, ou suffisamment de valeurs qui apparaissent statistiquement souvent dans l'ensemble de la population [1].

Voici un exemple de base de données anonyme qui vérifie le k-anonymat et la l-diversité (voir Figure 1.6). Ce que nous avons trouvé, c'est que nous avons en fait un meilleur anonymat ! Malheureusement, cela conduit inévitablement à une perte de précision [2] :

Nom	CP	Age	Poids
Sue	18000	22	50
Pat	69000	27	70
Bob	18500	21	90
Bill	18510	20	60
Dan	69100	26	70
Sam	69300	28	70

Données brutes

CP	Age	Poids
France	[20-29]	50
France	[20-29]	70
France	[20-29]	90
France	[20-29]	60
France	[20-29]	70
France	[20-29]	70

Données anonymes et diverses

FIGURE 1.6 – Tableaux données brutes & données anonymes et diverses [2].

Dans la Figure 1.6 ci-dessus, il ne nous reste qu'une seule catégorie, nous ne serons plus au niveau communal, ni même au niveau départemental, mais au niveau national. La tranche d'âge est plus large ici aussi [2].

Inconvénients

— Difficile à atteindre.

— Insuffisante pour empêcher la divulgation des attributs.

1.8.4 *t*-proximité

La *t*-proximité définit la "représentation" des valeurs, en forçant la distribution des données sensibles pour chaque classe d'équivalence à ressembler, au sein d'un facteur t, à la distribution générale de ces mêmes données sensibles. Puis il commence à se poser la question de l'utilité des données. Sous la contrainte de *t*-proximité, les données n'apparaissent pas nécessairement directement utilisables. Cependant, il est toujours possible d'identifier des tendances, de faire des calculs généraux ou des corrélations sur l'ensemble du tableau [1].

Comme on peut le voir sur la Table 1.2 suivante , la convergence t permet essentiellement de répondre à la question : comment diviser mes données de manière à ce que toutes les partitions soient similaires entre elles en termes de distribution ? Par exemple, si nous imaginons une base de données nationale sur les maladies, et comment les sections sont regroupées, groupes d'âge et sexe, de sorte que nous ayons la même répartition des maladies dans chaque sous-groupe. Nous pouvons poser des questions sur l'ensemble de données qui résulte de ce processus lorsque nous voulons spécifiquement effectuer une analyse qui met en évidence les facteurs qui caractérisent les individus [22].

Age	**Sexe**	**Département**	**Pathologie**	**Nombre d'individus**
<45	M	75	Grippe	400
<45	M	75	Rhume	800
>45	M	75	Grippe	500
>45	M	75	Rhume	1000
<35	F	75	Grippe	300
<35	F	75	Rhume	600
>35	F	75	Grippe	600
>35	F	75	Rhume	1200

TABLE 1.2 – t-proximité.

Inconvénients

— Détérioration de l'utilité des données.

— Plus la taille et la variété des données augmentent, plus les chances de ré-identification augmentent.

1.8.5 Confidentialité différentielle

La confidentialité différentielle [23] est une condition du mécanisme de publication des données et non de l'ensemble de données. Un algorithme randomisé est considéré comme différentiellement privé si, pour toute paire d'entrées voisines, la probabilité de générer la même sortie est inférieure à un petit multiple l'une de l'autre, pour l'ensemble de l'espace de sortie. Cela signifie que pour deux ensembles de données proches l'un de l'autre, un algorithme différentiellement privé se comportera approximativement de la même manière sur les deux ensembles de données. Cette notion offre une protection suffisante de la vie privée des utilisateurs, quelles que soient les connaissances préalables possédées par les adversaires. [24].

Cet algorithme satisfait les contraintes différentielles de confidentialité si [2] :

• Pour tout couple de tables (et donc pour les bases de données), D1 et D2 ne diffèrent que par la présence ou l'absence d'un individu.

• Pour tout résultat Ω de l'algorithme, il y a les éléments suivants :

$$Pr\left[\mathrm{A}\left(\mathrm{D}_1\right)=\Omega\right] \leq \mathrm{e}^{\varepsilon} Pr\left[\mathrm{A}\left(\mathrm{D}_2\right)=\Omega\right] \tag{1.1}$$

Inconvénients

— Les attaques par couplage d'enregistrements et le coût de calcul élevé.

1.9 Conclusion

Dans la première partie de ce chapitre nous avons présenté les concepts de vie privée et d'anonymat, nous avons ensuite présenté les méthodes qui permettent de protéger la vie privée ce qui peut être approprié, selon les circonstances et le contexte, pour atteindre l'objectif visé sans compromettre le droit des personnes concernées à la vie privée. A la fin de ce chapitre nous avons mis l'accent sur les différentes techniques de l'anonymat. Dans le chapitre suivant nous allons nous concentrer sur l'étude des méthodes de classification utilisées pour préserver et protéger l'anonymat des individus.

Chapitre 2

Classification et anonymat

2.1 Introduction

La classification est considérée comme un facteur clé pour la détection d'informations cachées dans les données volumineuses échangées. Les données stockées dans les bases de données contiennent souvent des informations sensibles, de sorte qu'une éventuelle divulgation pendant les processus d'extraction peut compromettre les droits fondamentaux des individus tels que la vie privée ou le droit de ne pas subir de discrimination. Dans ce chapitre, on va présenter les techniques de classification et quelques approches de classification qui ont été consacrées à la protection des informations sensibles.

2.2 La classification

La classification est une technique majeure d'exploration de données (data mining) et largement utilisée dans divers domaines. Il s'agit d'une technique d'exploration de données (apprentissage automatique) utilisée pour prédire l'appartenance à un groupe pour des états de données [25].

Il existe plusieurs grands types de méthodes de classification, y compris l'induction par arbre de décision, les réseaux bayésiens, le classificateur k-plus proche voisin.

2.3 Méthodes de classification

Il existe plusieurs types d'algorithmes de classification, chacun avec ses propres fonctionnalités et applications uniques. Tous ces algorithmes sont utilisés pour extraire des données de jeux de données.

2.3.1 L'arbre de décision

Les arbres de décision sont des arbres qui classent les instances en les triant selon les valeurs des paramètres [26].

L'algorithme de l'arbre de décision construit le modèle de classification sous la forme d'une structure arborescente. Utilise des règles conditionnelles complètes et contradictoires dans la classification. Le processus se poursuit en divisant les données en structures plus petites et en les fusionnant éventuellement avec un arbre de décision supplémentaire. La structure finale ressemble à un arbre avec des nœuds et des feuilles. Les règles sont apprises séquentiellement en utilisant les données d'entraînement une par une. Chaque fois qu'une règle est reconnue, les groupes qui couvrent les règles sont supprimés. Le processus se poursuit dans l'ensemble d'apprentissage jusqu'à ce que le point final soit atteint [25].
L'arbre de décision est composé de :

- **Noeuds de décision** : chacun contient un test sur l'attribut.
- **Branches** : correspondent généralement à l'une des valeurs possibles de l'attribut choisi.
- **Feuilles** : y compris les choses appartenant à la même classe.

Les étapes de construire l'arbre de décision à l'aide l'algorithme ID3 (par Ross Quinlan) [27] :

- Sélection un attribut pour le nœud racine : calculer l'entropie de Shanon.
- Supposons qu'il y ait deux catégories : (+) et (-). Considérons l'ensemble d'exemples S contenant p exemples de classe (+) et n exemples de classe (-).
- L'entropie est la quantité d'informations nécessaires pour décider si un exemple dans S appartient à (+) ou (-).

$$E(S) = -p_+ \log_2 (p_+) - p_- \log_2 (p_-) \tag{2.1}$$

Où, est la proportion des exemples (+), et est la proportion des exemples (-).

- Utilise le gain d'information pour mesurer la pureté des attributs.

$$Gain(S, A) = E(S) - \sum_{v \in valeur(A)} \frac{|S_v|}{|S|} E\left(S_v\right) \tag{2.2}$$

S_v représente le nombre d'instance totale de nœud S, et S représente le nombre d'instance totale de nœud parent.

- On choisit l'attribut qui a la plus grande valeur de Gain comme racine.
- Créez une branche pour chaque valeur possible de l'attribut [28].
  - Divisez les exemples en sous-ensemble, Un pour chaque branche partant du nœud.
  - Répétez de manière itérative pour chaque branche en utilisant uniquement les exemples liés à la branche.
  - Arrêter la récursivité d'une branche si tous les exemples ont la même classe.

Exemple

La Table 2.1 suivante représente un ensemble d'exemple pour jouer le Tennis.

Prévisions	Température	Humidité	Vent	Classe
Ensoleillé	Chaud	Haute	Faible	Non
Ensoleillé	Chaud	Haute	Fort	Non
Nuageux	Chaud	Haute	Faible	Oui
Pluvieux	Douce	Haute	Faible	Oui
Pluvieux	Fraîche	Normale	Faible	Oui
Pluvieux	Fraîche	Normale	Fort	Non
Nuageux	Fraîche	Normale	Fort	Oui
Ensoleillé	Douce	Haute	Faible	Non
Ensoleillé	fraîche	Normale	Faible	Oui
Pluvieux	Douce	Normale	Faible	Oui
Ensoleillé	Douce	Normale	Fort	Oui
Nuageux	Douce	Haute	Fort	Oui
Nuageux	Chaud	Normale	Faible	Oui
pluvieux	Douce	Haute	oui	Non

TABLE 2.1 – Ensemble d'exemple pour jouer le Tennis [10].

• Après avoir appliqué toutes les étapes de l'algorithme ID3 à l'ensemble du tableau précédent, on obtient cet arbre (voir Figure 2.1) :

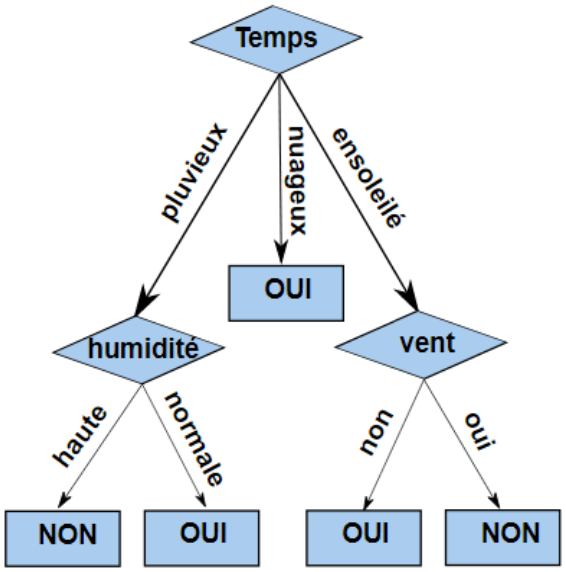

FIGURE 2.1 – L'arbre généré par ID3 de l'exemple jouer Tennis.

Il existe des autres algorithmes [10] :

— **Algorithme C4.5 (J48) :** c'est une amélioration d'ID3 qui prend en compte les attributs numériques et utilise le GainRation pour le Split/Segmentation.

— **Algorithme CART (Classification And Regression Trees) :** utilise l'indice de Gini pour le Split/Segmentation.

— **Forêts aléatoires (Random forest) :** plus efficaces mais difficilement interprétables, construction des arbres se base sur le Bootstrap (ou le Bagging).

2.3.2 Classification naïve bayésienne

Le classificateur Naïve Bayes est une méthode d'apprentissage bayésienne très pratique. il s'applique aux tâches d'apprentissage où chaque instance x est décrite par une conjonction de valeurs d'attributs et la fonction cible f(x) peut prendre n'importe quelle valeur d'un ensemble fini C. Un ensemble d'exemples d'apprentissage de la fonction cible est fourni, et une nouvelle instance est présentée, décrite par le tuple des valeurs d'attributs a_1, a_2, ..., a_n. L'apprenant est invité à prédire la valeur cible, ou la classification, pour cette nouvelle instance [29].

— Pourquoi bayésien ? Parce que l'algorithme est basé sur le théorème de Bayes qui nous dit que [30] :

$$P(A \mid B) = P(B \mid A)^{*}P(A)/P(B) \tag{2.3}$$

— Pourquoi naïf ? Puisque cet algorithme est utilisé, nous supposons que les variables explicatives sont indépendantes, ce qui est en fait faux. Mais nous l'acceptons quand même pour pouvoir l'utiliser [30].

La classification Naive Bayes est décrit par l'expression suivante :

$$P\left(C_i \mid X\right) = \frac{P\left(X \mid C_i\right) * P\left(C_i\right)}{P(X)} \tag{2.4}$$

$$D'ou$$

$$P\left(X \mid C_i\right) = \prod_{k=1}^{n} P\left(X_K \mid C_i\right) \tag{2.5}$$

Exemple

Voici un set (Table 2.2) qu'on va utiliser pour prédire la classe de X (a2, b1, c3, d1, ?) [11] :

	A	B	C	D	Classe
E1	a1	b1	c1	d2	+
E2	a1	b2	c2	d2	+
E3	a1	b2	c3	d1	-
E4	a2	b1	c1	d1	-
E5	a2	b2	c1	d1	-
E6	a2	b2	c1	d2	+
E7	a1	b1	c1	d1	+
E8	a2	b1	c2	d2	-
E9	a3	b1	c3	d1	+
E10	a3	b2	c2	d2	+

TABLE 2.2 – Exemple d'une data set.

— La probabilité que X est classifié dans la classe (+) :
$P(+ \mid X) = (P(a2 \mid +) * P(b1 \mid +) * P(c3 \mid +) * P(d1 \mid +) * P(+))/\ P(a2) * P(b1) * P(c3) * P(d1) = (1/6 * 3/6 * 1/6 * 2/6 * 6/10)/(4/10 * 5/10 * 2/10 * 5/10) = 0,0027/P(X)$

— La probabilité que X est classifié dans la classe (-) :
$P(- \mid X) = (P(a2 \mid -) * P(b1 \mid -) * P(c3 \mid -) * P(d1 \mid -) * P(-))/\ P(a2) * P(b1) * P(c3) * P(d1) = (3/4 * 2/4 * 1/4 * 3/4 * 4/10)/(4/10 * 5/10 * 2/10 * 5/10) = 0,02/P(X)$

Donc X est classifié dans la classe (-).

Il existe plusieurs types de classificateur Naïve Bayes tels que :

— Gaussian Naïve Bayes

— Multinomial Naïve Bayes

— Bernoulli Naïve Bayes

— Complement Naïve Bayes

— Categorical Naïve Bayes

2.3.3 Support vecteur machine (SVM)

Les machines à vecteurs de support sont des modèles d'apprentissage supervisés avec des algorithmes d'apprentissage associés qui analysent les données et reconnaissent les modèles. Le SVM de base prend un ensemble de données d'entrée et prédit, pour chaque entrée donnée, laquelle des deux classes possibles forme la sortie, ce qui en fait un classificateur linéaire binaire non probabiliste. Les modèles SVM ont une forme fonctionnelle similaire aux réseaux de neurones, leur principe est simple : ils visent à séparer les données en classes en utilisant le "simple" maximum possible, c'est-à-dire la distance entre les différents ensembles de données et le maximum entre eux [31].

L'hyperplan séparateur est représenté par l'équation :

$$H(X) = w^T x + b \tag{2.6}$$

W est un vecteur de dimensions m et b est un terme. La fonction de décision, pour un exemple x, peut être exprimée comme suit :

$$\begin{cases} Classe = 1 \;\; Si \quad H(x) > 1 \\ Classe = -1 \;\; Si \quad H(x) < -1 \end{cases} \tag{2.7}$$

Maximiser la marge revient maximiser $\frac{2}{\|w\|}$ et qui veut à minimiser $\|w\|$.

La figure 2.2 représente l'illustration de l'hyperplan séparateur et la marge maximale.

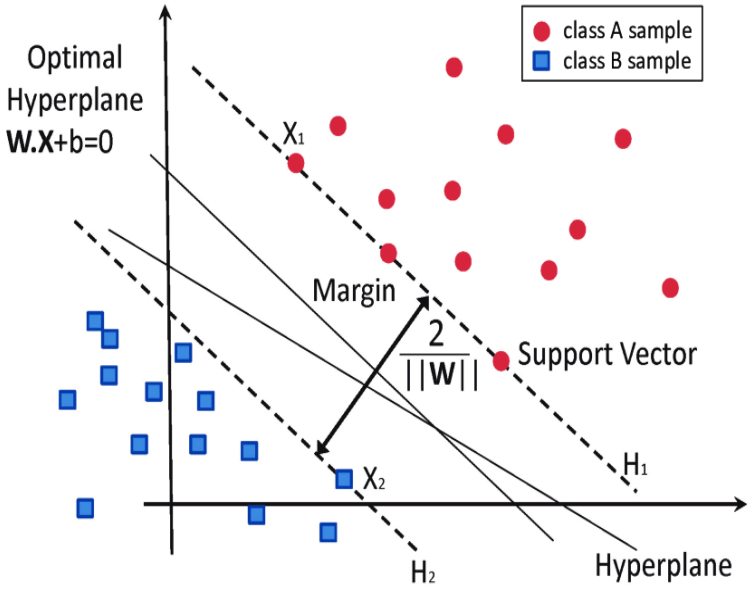

FIGURE 2.2 – Illustration de l'hyperplan séparateur et la marge maximale [3].

Exemple

La Table 2.3 suivante représente un exemple d'une data set [4] :

X1	**X2**	**Classe**
1	yes	-1
2	no	-1
3	no	-1
4	yes	-1
5	no	-1
6	no	1
7	yes	1
8	no	1
9	yes	1
10	yes	1
11	yes	1
12	yes	1
13	yes	1

TABLE 2.3 – Exemple d'une Data set.

On vas représenter les points dans le data set sur un plan (x,y) et extraire les vecteurs support pour chaque classe (voir Figure 2.3)[4].

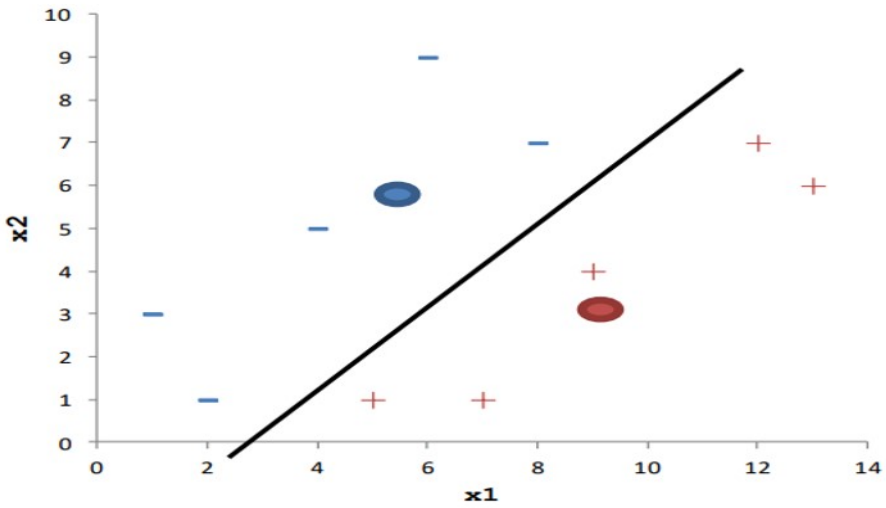

FIGURE 2.3 – Data set avec support vector machine [4].

2.3.4 K nearest neighbors

La méthode k plus proches voisins est l'une des méthodes d'apprentissage supervisé les plus simples qui peuvent être utilisées pour les régressions et les classifications. Elle a pour but de calculer la distance entre tous les exemples de la règle et le nouvel exemple que l'on cherche à classer et de choisir la classe majoritaire parmi les K distances les plus petites [32].

Pour appliquer cette méthode, la marche à suivre est la suivante [32] :

- On fixe le nombre de voisins k.
- Nous découvrons les k voisins les plus proches des nouvelles données d'entrée que nous voulons classer.
- Les catégories correspondantes sont attribuées à la majorité des voix.

Mais comment choisit-on ce paramètre k lors de l'exécution de l'algorithme ? [32]

- Nous différons k.
- Pour chaque valeur de k, nous calculons le taux d'erreur de l'ensemble de test
- Nous maintenons le paramètre k qui réduit ce taux d'erreur de test.

Les métriques les plus souvent choisies sont la distance usuelle dite euclidienne et Manhattan :

— La distance euclidienne :

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{2.8}$$

— La distance Manhattan :

$$\sum_{i=1}^{k}|x_i - y_i| \tag{2.9}$$

Exemple

On va appliquer la distance euclidienne pour choisir la classe majoritaire parmi les k-distance les plus petits (voir table 2.4) :

On pose k=3 et le test data est comme suit :

Scénario = 3, jeu d'acteur = 7, classe = ?

	Scénario	**Jeu d'acteur**	**Classe**	**Distance**
1	7	7	bon	sqrt[(7-3)^2+(7-7)^2]=4
2	7	4	bon	sqrt[(7-3)^2+(4-7)^2]=5
3	3	4	mauvais	3
4	1	4	mauvais	3.60

TABLE 2.4 – Exemple d'une Data set [11].

Les trois plus petites distances sont : 3, 4 et 1 qui ont classé respectivement comme suit : mauvais, mauvais et bon. Donc la classe de notre test data est « mauvais » [11].

2.3.5 Les réseaux de neurones artificiels

Les RNA consistent en une couche de nœuds d'entrée et une couche de nœuds de sortie, reliées par une ou plusieurs couches de nœuds cachés. Les nœuds de couche d'entrée transmettent des informations aux nœuds de couche cachés en déclenchant des fonctions d'activation, et les nœuds de couche cachés se déclenchent ou restent inactifs en fonction des preuves présentées. Les couches cachées appliquent des fonctions de pondération aux preuves, et lorsque la valeur d'un nœud particulier ou d'un ensemble de nœuds dans la couche cachée atteint un certain seuil, une valeur est transmise à un ou plusieurs nœuds

dans la couche de sortie. Les RNA doivent être formés avec un grand nombre de cas (données) [33].

Exemple

Ici on veut déduire la préférence politique du futur électeur, on aura déjà pas mal d'informations [Table 2.5], on utilise un réseau de neurones pour ceux à qui il manque un critère [5].

Age	Revenue	Education	Sexe	Politique
24	24000,00 €	Bas	Féminin	Gauche
62	82000,00 €	Moyen	Masculin	Droite
38	64000,00 €	Haut	Féminin	Autre
30	40000,00 €	Moyen	Masculin	Droite
45	42000,00 €	Haut	Féminin	Gauche
35	49000,00 €	Haut	Masculin	A prédire

TABLE 2.5 – Exemple d'une Data set.

Donc nous voudrions définir le parti politique d'un homme de 35 ans qui gagne 49000 € avec un haut niveau d'éducation [5].

Après la phase de normalisation, la création de notre réseau de neurones, et un apprentissage réussi, voici à quoi cela pourrait ressembler [Figure 2.4] [5] :

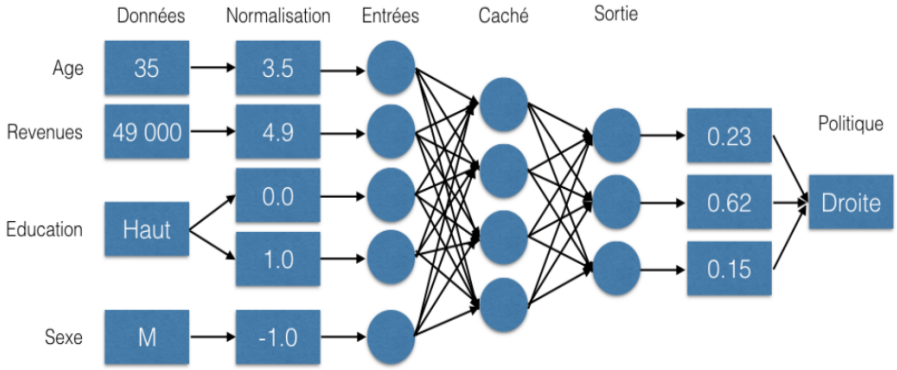

FIGURE 2.4 – Prédiction d'un critère avec réseau de neurones [5].

2.4 Anonymat pour la classification

2.4.1 La distribution des données

La nature de la distribution des données (centralisée ou distribuée) est un facteur important à prendre en compte lors de la conception de solutions de confidentialité [34].

Données centralisées

Les données centralisées sont des métadonnées qui sont collectés sur un serveur/plateforme dédié.

La figure 2.5 représente un exemple d'un système des données centralisées.

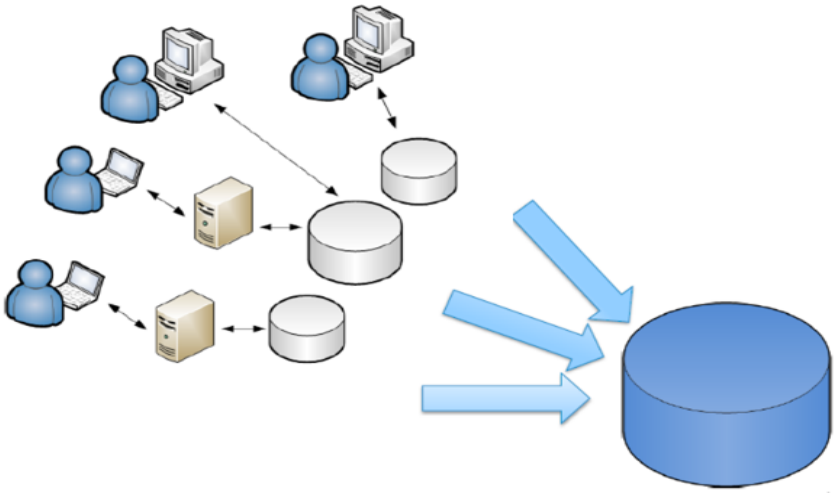

FIGURE 2.5 – Exemple d'un système de données centralisées [6].

Données distribuées

Les données distribuées sont des données fragmentées ou répliquées sur les différentes configurations matérielles et logicielles, situées généralement sur différents sites géographiques au sein d'une organisation [7].

La figure 2.6 représente un exemple d'un système des données distribuées.

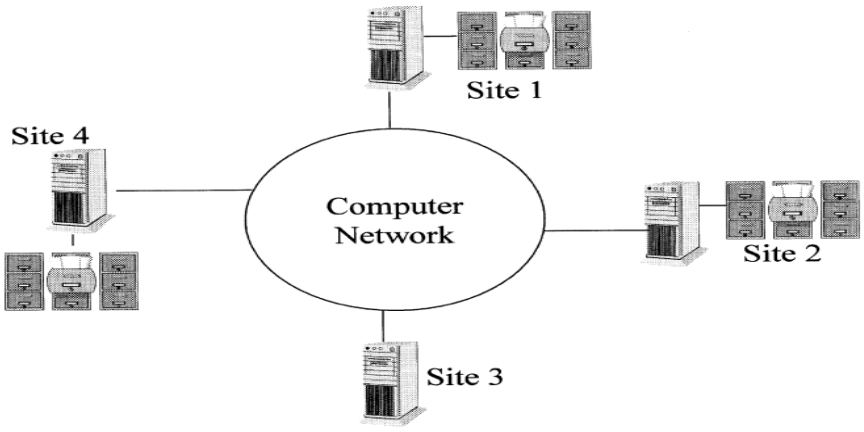

FIGURE 2.6 – Exemple d'un système des données distribuées [7].

— **Distribution horizontale :** chaque site contient une partie des éditions de la base de données mondiale. Tous les sites partagent le même schéma (tous les attributs, y compris l'attribut de catégorie dans le cas de la classification modérée) (voir Figure 2.7).

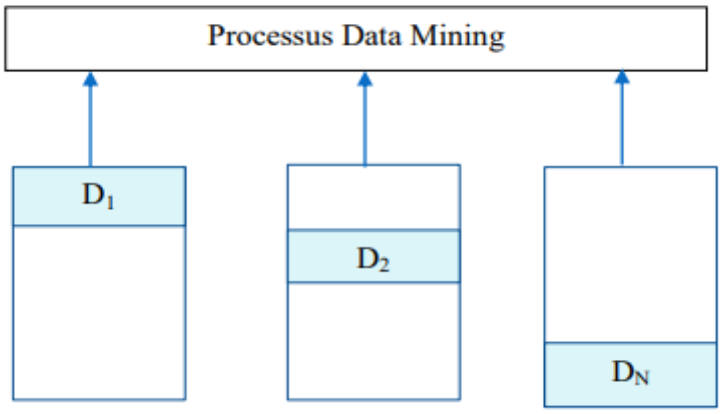

FIGURE 2.7 – Distribution horizontale des données [8].

— **Distribution verticale :** chaque site contient un sous-ensemble de traits mais avec un schéma différent (traits différents).

La figure 2.8 suivante montrent la distribution verticale des données.

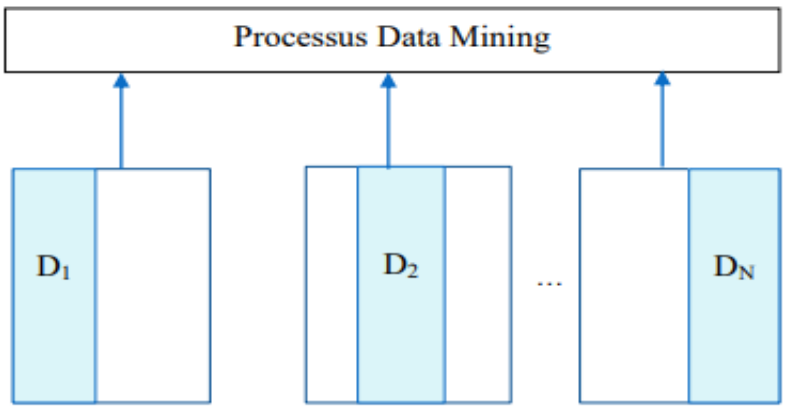

FIGURE 2.8 – Distribution verticale des données [8].

2.4.2 Les approches de classification préservant l'anonymat

L'extraction de données est largement utilisée par les chercheurs à des fins scientifiques et commerciales. Les données recueillies auprès des individus sont uniques pour le prix de décision ou la reconnaissance de modèles. Par conséquent, des processus de préservation de la vie privée ont été développés pour assainir les informations privées des échantillons tout en conservant leur utilité. Ci-dessous, nous expliquons certains d'entre eux.

La figure 2.9 suivante représente les approches de classification préservant l'anonymat.

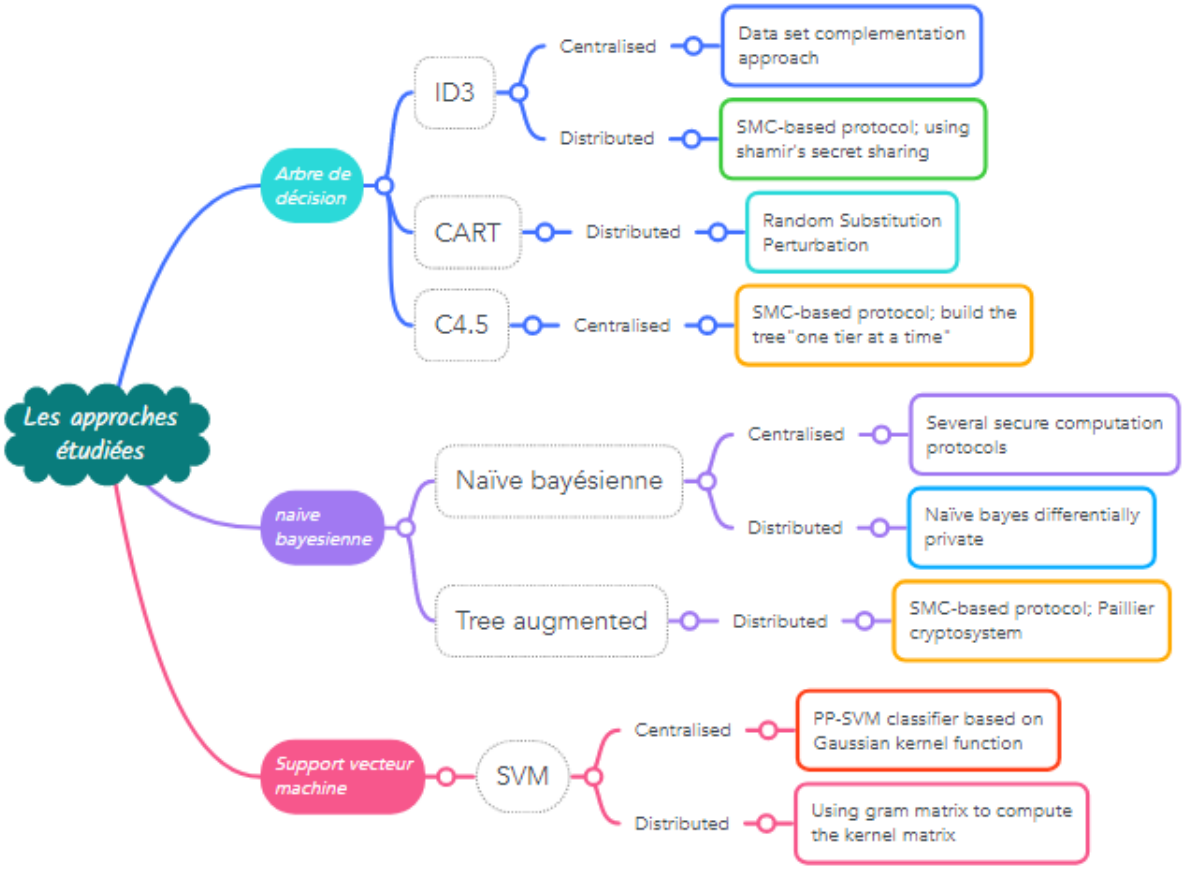

FIGURE 2.9 – Les approches de classification préservant l'anonymat.

A. Arbre de décision

A.1. Random Substitution Perturbation

L'idée de base est de remplacer la valeur de chaque enregistrement de données sous l'attribut par une autre valeur choisie aléatoirement dans le domaine de l'attribut selon un modèle probabiliste. Dowd et d'autres montrent qu'une telle perturbation est immunisée contre l'attaque de récupération de données qui vise à récupérer les données d'origine à partir des données perturbées, et l'attaque de perturbation répétée où un adversaire peut perturber à plusieurs reprises les données dans l'espoir de récupérer les données d'origine

[35] [36].

Les étapes de réalisation

— Génération d'un algorithme de perturbation par substitution aléatoire et analyser son immunité à l'attaque de récupération de données.
— Génération d'un l'algorithme de reconstruction avec des méthodes heuristiques pour réduire l'erreur d'estimation des distributions de données originales.
— Analyser l'effet des paramètres de la matrice de perturbation et l'immunité à l'attaque de perturbations répétées.
— Présentation des résultats des expériences.
— Discution de la manière de sélectionner les paramètres de la matrice de perturbation dans la pratique.

A.2. SMC-based protocol (build the tree "one tier at a time")

Brickell et Shmatikov présentent un protocole cryptographiquement sécurisé pour construire des arbres de décision préservant la confidentialité. Le protocole a lieu entre un utilisateur et un serveur. L'entrée de l'utilisateur se compose des paramètres de l'arbre de décision qu'il souhaite créer, tels que les attributs qui sont traités comme des fonctionnalités et l'attribut qui représente la classe. L'entrée du serveur est une BDD relationnelle. La sortie du protocole utilisateur est un arbre de décision généré à partir des données du serveur, tandis que le serveur n'apprend rien de l'arbre généré [37].

Les étapes de réalisation

— Partage des valeurs d'attribut.
— Calcul du nombre de catégories.
— Sélection du fractionnement de la plus haute qualité.
— Construire le niveau inférieur.

A.3. Data set complementation approach

Fong et d'autres, introduisent une approche basée sur la perturbation et la randomisation pour protéger les ensembles de données utilisés dans l'exploration d'arbres de décision. Avant d'être transmis à un tiers pour la construction d'un arbre de décision, les ensembles de données d'origine sont convertis en un groupe d'ensembles de données irréels, à partir desquels les données d'origine ne peuvent pas être reconstruites sans l'ensemble du groupe d'ensembles de données irréelles. Pendant ce temps, un arbre de décision précis peut être

construit directement à partir des ensembles de données irréels [38][9].

Les étapes de réalisation

- Définition l'ensemble universel et l'ensemble de données complémentaires.
- Génèration des ensembles de données irréalisés qui sont utilisés pour convertir les jeux de données en un ensemble d'apprentissage irréalisé.
- Génération l'arbre de décision avec l'algorithme ID3 en appelant récursivement l'algorithme Choose-Attribute.
- Déterminer Les entropies des jeux de données d'origine, avec tout attribut de décision et tout attribut de test, par l'ensemble d'apprentissage irréel et l'ensemble de perturbation.
- Introduire un algorithme d'apprentissage d'arbre de décision modifié utilisant l'ensemble d'apprentissage irréalisé et l'ensemble de perturbation.
- Reconstruire les jeux de données d'échantillons originaux, à partir de l'ensemble d'apprentissage et de l'ensemble de perturbation.
- Evaluation théorique de l'approche.

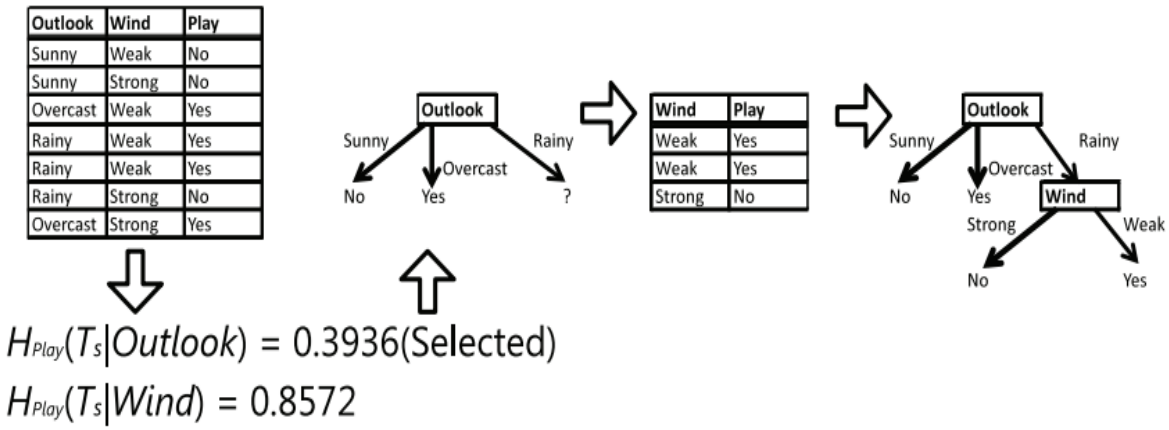

FIGURE 2.10 – Illustration en appliquant l'approche ID3 classique avec les échantillons originaux T_S [9].

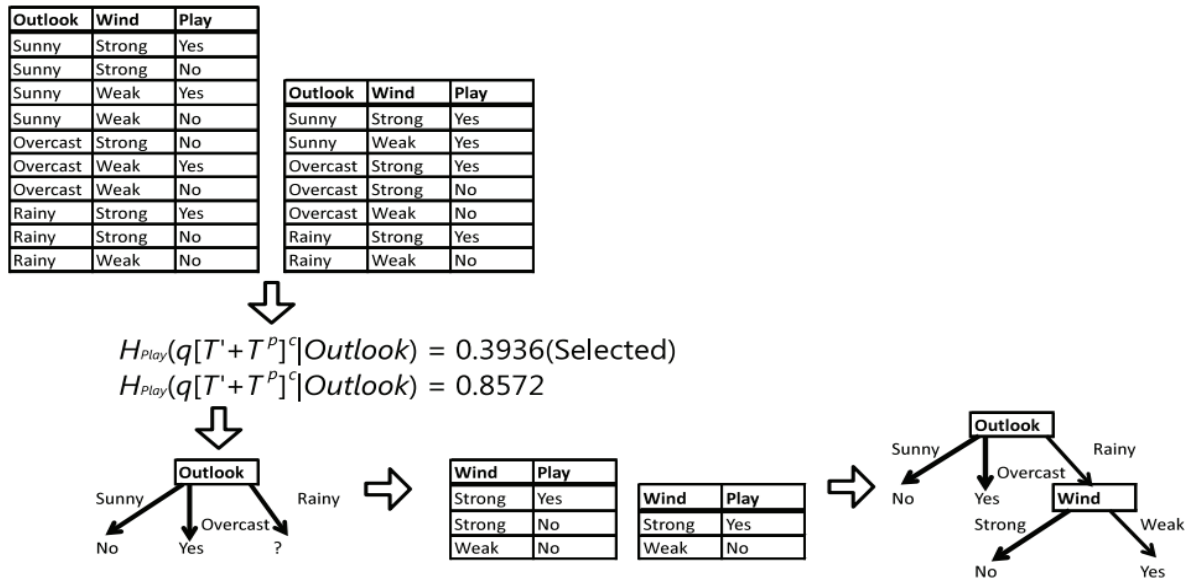

FIGURE 2.11 – llustration en appliquant l'approche ID3 modifiée avec les échantillons irréalisés ($T^{'} + T^{P}$). Pour chaque étape, les valeurs d'entropie et les sous-arbres résultants sont exactement les mêmes que ceux de l'approche traditionnelle [9].

A.4. SMC-based protocol (using shamir's secret sharing)

Sheela et Vijayalakshmi proposent une méthode basée sur SMC pour construire un arbre de décision préservant la confidentialité sur des données segmentées verticalement. La méthode proposée utilise l'algorithme de partage Shamir Secret pour calculer en toute sécurité la relation de base du produit ponctuel, qui est nécessaire lors du calcul du gain d'information des attributs lors de la construction de l'arbre de décision [39].

Les étapes de réalisation

- Introduction de L'algorithme d'arbre de décision ID3 qui recherche les attributs et produisent le maximum d'informations pour déterminer l'appartenance à une classe d'instances dans un ensemble d'apprentissage d'instances étiquetées.
- Cardinalité sécurisée du produit scalaire en utilisant le Shamir's secret sharing (Algorithme de Shamir's secret sharing à chaque partie, recherche d'une somme sécurisée à l'aide de Shamir's share, méthode proposée pour trouver en toute sécurité la cardinalité du produit scalaire).
- Analyse de l'algorithme.

Secure Multi-party Computation(SMC)

Lorsque deux ou plusieurs parties souhaitent effectuer l'exploration de leurs données sans révéler leurs données privées mais en ne révélant que leur sortie finale, cela s'appelle un problème de calcul multipartite sécurisé (SMC). Les méthodes sécurisées basées sur le calcul multipartite peuvent prendre en charge l'exploration de données préservant la confidentialité [40]. Les méthodes décrites incluent la somme sécurisée, l'union d'ensembles sécurisés, la taille sécurisée de l'intersection d'ensembles et le produit scalaire. Une évaluation polynomiale sécurisée à deux parties a été effectuée à l'aide d'un produit scalaire vectoriel [41][39].

B. Naïve Bayésienne

B.1. Several secure computation protocols

Sur la base d'études antérieures sur le calcul multipartite sécurisé, Vaidya et d'autres proposent différents protocoles pour apprendre des modèles de classification bayésiens naïfs à partir de données partitionnées verticalement ou horizontalement. Pour les données partitionnées horizontalement, tous les attributs nécessaires à la classification d'une instance sont détenus par un seul site. Chaque partie peut obtenir directement le résultat de la classification, il n'est donc pas nécessaire de masquer le modèle de classification. Alors que pour les données partitionnées verticalement, puisqu'une partie ne connaît pas tous les attributs de l'instance, elle ne peut pas apprendre le modèle complet, ce qui signifie que le partage du modèle de classification est nécessaire. Dans ce cas, des protocoles qui peuvent empêcher la divulgation d'informations sensibles contenues dans le modèle de classification (par exemple, des distributions d'attributs sensibles) sont souhaités [29] [42].

Les étapes de réalisation

1. **Naïve Bayes préservant la confidentialité pour les données partitionnées horizontalement :**
   - Construction du modèle de classificateur pour le calcul des paramètres(attributs nominaux/numériques).
   - Evaluation de classificateur.

— Amélioration de la sécurité en utilisant une approche basé sur le logarithme sécurisé.

2. **Naïve Bayes préservant la confidentialité pour les données partitionnées verticalement :**
   — Construction du modèle de classificateur pour le calcul des paramètres(attributs nominaux/numériques).
   — Evaluation de classificateur.

B.2. SMC-based protocol (paillier cryptosystem)

Skarkala et d'autres, proposent la version préservant la confidentialité du classificateur Tree Augmented Naïve Bayesian utilisé par un protocole qui vise à extraire des informations globales à partir de bases de données statistiques partitionnées horizontalement. Le protocole présenté a été développé dans un environnement client-serveur et les participants ne peuvent être connectés qu'avec le mineur, ce qui rend la communication entre eux impossible, où La confidentialité est préservée à l'aide de techniques cryptographiques exploitant les primitives homomorphes [43].

Paillier cryptosystème

Le cryptosystème Paillier est basé sur le problème que le calcul des nièmes classes de résidus est un calcul intensif. La nature de l'algorithme permet aux opérations d'addition homomorphe de produire la réponse actuelle une fois déchiffrée [43].

Tree Augmented Naïve Bayesian (TAN)

TAN est un réseau bayésien à structure en forme de talon, qui est l'extension naturelle du modèle NaiveBayesian. Son idée principale est de combiner la simplicité du classificateur bayésien naïf avec la capacité d'exprimer la dépendance entre les attributs dans le réseau bayésien, ce qui peut améliorer les performances de classification du classificateur [44].

Les étapes de réalisation

— Générateur de clé qui comprend la génération de la paire de clés de chiffrement et la création de la signature numérique.
— Connexion et authentification mutuelle et envoyer les données partitionnée horizontalement.
— Trois clients sont impliqués pour initialiser le classificateur naive bayes.

— Collecter les fréquences pour chaque attribut de tous les clients.

— Création du modèle TAN.

— Envoyer les résultats finaux à tous les clients.

B.3. Naïve bayes differentially private

Vaidya et d'autres envisagent un scénario centralisé, où le mineur de données a un accès centralisé à un ensemble de données. Le mineur souhaite publier un classificateur en partant du principe que les informations sensibles sur les propriétaires de données d'origine ne peuvent pas être déduites du modèle de classification. Ils utilisent un modèle de confidentialité différentiel pour construire un classificateur bayésien naïf préservant la confidentialité. L'idée de base est de dériver la sensibilité pour chaque attribut et d'utiliser la sensibilité pour calculer le bruit laplacien. En ajoutant du bruit aux paramètres du classificateur, le mineur de données peut obtenir un classificateur qui est garanti différentiellement privé [24].

Le mécanisme laplacien

Le mécanisme laplacien est le mécanisme le plus courant pour assurer le secret différentiel, qui fonctionne en ajoutant du bruit aléatoire en réponse à une requête. Tout d'abord, la valeur réelle de f(D) est calculée, où f est la fonction de requête et D est l'ensemble de données, puis un bruit aléatoire est ajouté à f(D) et la réponse est renvoyée A(D) = f(D) + bruit est finalement retournée. L'amplitude du bruit est choisie en fonction du plus grand changement que l'enregistrement peut provoquer dans la sortie de la fonction de requête (par exemple, cela peut correspondre à 1 pour une requête à calculer via D1 et D2) [45].

Les étapes de réalisation

— Dériver la sensibilité de chaque attribut de manière appropriée selon qu'il est catégoriel ou numérique.

— Ajouter le bruit laplacien de l'échelle appropriée (et moyenne 0) aux paramètres (les comptages pour les attributs catégoriques, les moyennes et les écarts-types pour les attributs numériques).

— Utiliser les paramètres calculés pour classer une nouvelle instance à la manière standard de Naïve Bayes.

C. Support vecteur machine

B.1 Using gram matrix to compute the kernel matrix

Vaidya et d'autres, proposent une solution pour construire un modèle de classification global SVM à partir de données distribuées à plusieurs parties, sans divulguer les données de chaque partie. Ils considèrent la matrice noyau, qui est la structure centrale d'une SVM, comme un profil intermédiaire qui ne divulgue aucune information sur les données locales mais peut générer le modèle global. Ils proposent une méthode basée sur le calcul de la matrice gram pour calculer en toute sécurité la matrice du noyau à partir des données distribuées [46].

Noyau (kernel)

La fonction kernel (noyau) est une méthode utilisée pour prendre des données en entrée et les transformer en la forme requise de traitement des données. "Kernel" est utilisé en raison d'un ensemble de fonctions mathématiques utilisées dans Support Vector Machine fournissant la fenêtre pour manipuler les données [47].

Matrice de Gram (gram matrix)

La matrice de Gram (gram matrix) est simplement la matrice du produit interne de chaque vecteur et de ses vecteurs correspondants [48].

La matrice du noyau (kernel matrix)

La matrice du noyau (kernel matrix) est la structure centrale d'une SVM. Il contient toutes les informations nécessaires à l'algorithme d'apprentissage et fusionne les informations sur les données et le noyau [49].

Les étapes de réalisation

1. **Développement d'une technique PP-SVM pour les données partitionnées verticalement :**
   - Collecter des informations différentes sur le même ensemble d'entités.
   - Construire la matrice gram globale à partir des matrices gram locales afin que chaque partie puisse exécuter un solveur de programmation quadratique pour calculer le modèle SVM global.
   - Décrivent une méthode simple pour calculer en toute sécurité la somme des nombres entiers de sites individuels en supposant qu'il y a au moins trois parties et que les parties ne s'entendent pas.

— Etendre la méthode afin de fusionner de manière transparente les modèles locaux avec une efficacité et une confidentialité élevées.
— Calculer des produits scalaires entre les vecteurs de support et le nouvel objet de données pour tester un nouvel objet de données à l'aide du modèle.
— Evaluation.

2. **Esquisser brièvement la solution pour les données partitionnées horizontalement.**
3. **Développement de la solution pour les données arbitrairement partitionnées :**
— générer la matrice gram, un élément à la fois, en utilisant une version modifiée du protocole de produit scalaire basé sur le cryptage homomorphe.
— modification de l'algorithme pour le rendre plus résistant aux collusions.

B.2 PP-SVM classifier based on Gaussian kernel function

Lin et Chen proposent un classificateur SVM préservant la confidentialité basé sur la fonction de kernel gaussien [50]. La préservation de la vie privée est réalisée en transformant la fonction de décision d'origine, qui est déterminée par des vecteurs de support, en une série infinie de combinaisons linéaires de vecteurs de support mappés de caractéristiques monomiales. Le contenu sensible des vecteurs de support est détruit par la combinaison linéaire, tandis que la fonction de décision peut se rapprocher précisément de celle d'origine [51].

Noyau gaussien (gaussian kernel)

Le noyau gaussien transforme le produit scalaire dans l'espace dimensionnel infini en fonction gaussienne de la distance entre les points dans l'espace des données : si deux points dans l'espace des données sont proches, l'angle entre les vecteurs qui les représentent dans l'espace du noyau sera petit [52].

Les étapes de réalisation

— Construction de la fonction de décision préservant la vie privée à l'aide de cartographie des caractéristiques monômes.
— Approximation de la fonction de décision préservant la confidentialité.
— Discuter les problèmes de sécurité et de précision d'approximation du PPSVC.

2.4.3 Étude comparative des techniques precédente

Techniques	**Critères**	**Evaluation**
Random Substitution Perturbation [36]	-Erreur d'estimation de la distribution des données	-Réduire l'erreur d'estimation de 50 % à l'aide de l'heuristique.
	-Précision de la classification	-Très différente de celle de jeux de données d'origine.
SMC-based protocol (build the tree"one tier at a time") [37]	-Temps de connexion requis par le protocole	-Dépend de : facteur de branchement, nombre d'attributs de caractéristiques, nombre de niveaux et le nombre d'enregistrements.
Data set complementation approach [38]	-Précision de la classification	-Les arbres de décision générés sont les mêmes que les arbres de décision originaux.
	-Complexité du stockage	-Les exigences de stockage est bien inférieure à celle des échantillons d'origine.
	-Perte de confidentialité	-Elimine le risque de la confidentialité et améliore la sécurité.
SMC-based protocol (using shamir's secret sharing) [39]	-Effet de la collusion sur la sécurité	-Utilise une méthode sécurisée efficace pour trouver la cardinalité du produit scalaire.
	-Coût des communications	-Utilise moins de coût de communication.
	-Coût de calcul	-Utilise moins de coût de calcul.
SMC-based protocol (paillier cryptosystem) [43]	-Temps de calcul	-Non seulement efficace mais aussi efficient.
	-Précision de la classification	-Meilleure précision par rapport au modèle Naïve Bayes.

Several secure computation protocols [29]	-Effet de la collusion sur la sécurité	-Parties en collusion peuvent apprendre les valeurs que nous voulons protéger (horizontal). -Résistant à la collusion (vertical).
	-Coût des communications	-la version sécurisée est nettement plus lente (horizontal). -Le temps nécessaire reste tout à fait raisonnable (vertical).
	-Coût de calcul	-la version sécurisée est nettement plus lente (horizontal). -Le temps nécessaire reste tout à fait raisonnable (vertical).
Naïve bayes differentially private [24]	-Précision de la classification	-Fonctionne très bien et capable de suivre le classificateur Naïve Bayes de base même tout en offrant une très forte confidentialité.
Using gram matrix to compute the kernel matrix [46]	-Effet de la collusion sur la sécurité	-Préserve la confidentialité des données(vertical). -Résistant à la collision (arbitraire).
	-Coût de calcul	-Ne change guère (vertical). -Assez raisonnable (arbitraire).
	-Coût des communications	-N'est pas visible (vertical). -Assez raisonnable (arbitraire).
PP-SVM classifier based on Gaussian kernel function [51]	-Sécurité contre les attaques sur les vecteurs supports	-L'attaquant qui ne connaît qu'une partie des données d'entraînement ne peut pas obtenir les supports.
	-Approximation de la précision	-Le PPSVC peut approcher avec précision le classificateur SVM d'origine par un faible degré d'approximation.

TABLE 2.6 – Étude comparative des techniques precédentes.

2.5 Conclusion

Dans ce chapitre, nous avons présenté l'une des techniques suppervisée de fouille de données "la classification" et ses méthodes de manière générale, ainsi que l'étude des techniques de classification utilisées pour preserver et protéger l'anonymat des individus. Nous avons proposé une carte mentale pour classer les différentes méthodes utilisés selon leur mode de fonctionnement. Ensuite nous avons terminé cette partie avec un tableau comparatif qui permet de faire une comparaison entre les différentes techniques et les critères d'évaluation de chacune.

Dans le prochain chapitre, nous allons présenter notre propre méthode pour préserver l'anonymat des individus dans le cas de la classification, et l'évaluation de cette solution.

Chapitre 3

Approche proposée et validation

3.1 Introduction

Dans ce dernier chapitre, nous allons présenter d'abord c'est quoi la problèmatique de notre thème et la solution que nous avons proposé pour la résoudre, puis nous présenterons notre dataset avec une description de ses caractéristiques et les étapes de pré-traitement des données que nous avons effectué sur ce dataset, ainsi que nous définissons les différentes bibliothèques et langage de développement utilisés pour le processus d'implémentation et d'évaluation.

3.2 Problématique

La préservation de la vie privée est importante pour l'apprentissage automatique et l'exploration de données, mais les mesures conçues pour protéger les informations privées entraînent souvent un compromis : une utilité réduite des échantillons d'apprentissage, pour cela, nous introduisons une approche de préservation de l'anonymat qui peut être appliquée à la classification naive bayesienne GaussianNB, sans perte de précision et d'utilité des données.

La figure 3.1 représente l'objectif de notre travail.

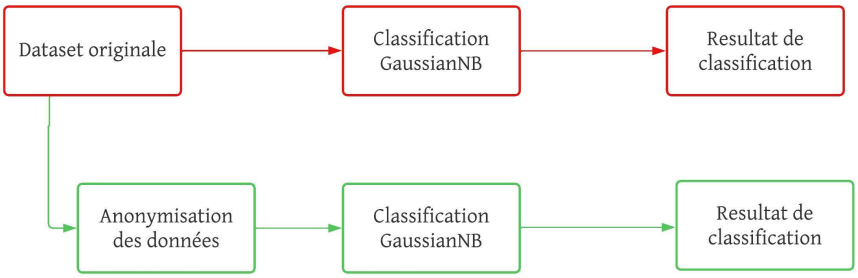

FIGURE 3.1 – Objectif de notre travail.

3.3 Architecture de notre approche

Après avoir soulevé notre problématique dans la section précedente, maintenant nous allons présenter une idée générale de l'approche qu'on a proposé et qui consiste à prturber le jeu de données en ajoutant un nombre limité de fausses lignes afin d'anonymiser le dataset tout en conservant la qualité des informations.

La figure 3.2 représente l'architecture générale de notre proposition.

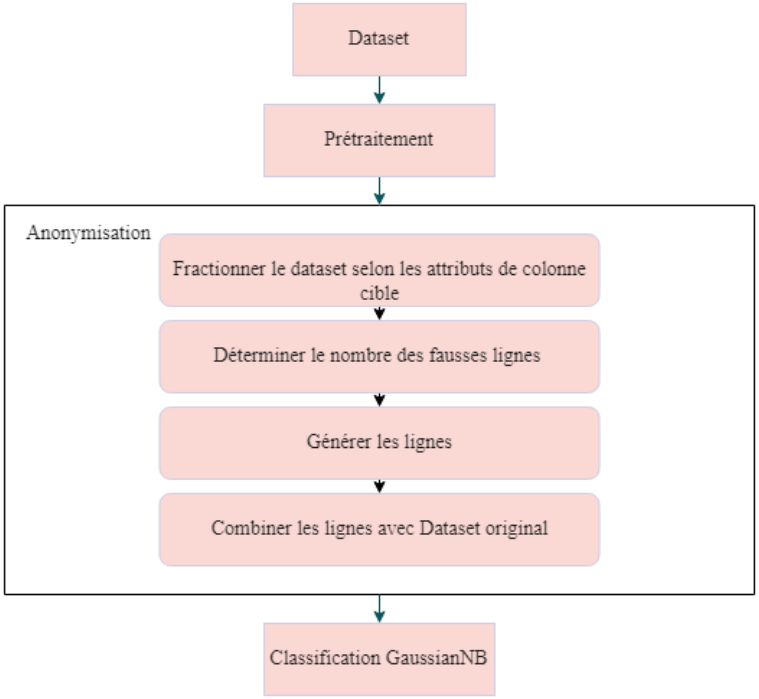

FIGURE 3.2 – Architecture générale de notre proposition.

3.4 Nettoyage et pré-traitement des données

Lorsqu'il s'agit de créer un modèle de Machine Learning, le prétraitement des données est la première étape marquant l'initiation du processus. En règle générale, les données du monde réel sont incomplètes, incohérentes, inexactes (contiennent des erreurs ou des valeurs aberrantes) et manquent souvent de valeurs/tendances d'attributs spécifiques. C'est là que le prétraitement des données entre dans le scénario ; il aide à nettoyer, formater et organiser les données brutes, les rendant ainsi prêtes à l'emploi pour les modèles d'apprentissage automatique, nous explorerons notre prétraitement des données dans l'apprentissage automatique.

3.4.1 Création du jeu de données(La matrice de données)

Dans notre étude, nous allons utiliser le jeu de données **Heart Attack Analysis** & **Prediction Dataset** [53] qui continent deux fichiers au format csv :

— Le fichier "heart" comprend une description détaillé d'analyse et de prévision des crises cardiaques.

— Le fichier "O_2 Saturation" comprend une description sur le niveau de saturation.

Dans notre travail nous allons travailler sur le dataset Heart Attack Analysis & Prediction Dataset, exactement le fichier "heart" qui contient 303 lignes et 14 colonnes (voir figure 3.3).

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

FIGURE 3.3 – Dataset avant prétraitement.

3.4.2 La description des variables de fichier "heart"

- **Age :** age du patient.
- **Sex :** sexe du patient.
- **exang :** angine induite par l'effort (1 = oui ; 0 = non).
- **ca :** nombre de navires principaux (0-3).
- **cp :** type de douleur thoracique.
  - **Value 1 :** angine typique.
  - **Value 2 :** angine atypique.
  - **Value 3 :** douleur non angineuse.
  - **Value 4 :** asymptomatique.
- **trtbps :** tension artérielle au repos (en mm Hg).
- **chol :** cholestoral en mg/dl récupéré via le capteur IMC.
- **fbs :** (glycémie à jeun > 120 mg/dl) (1 = vrai ; 0 = faux).
- **rest_ecg :** résultats électrocardiographiques au repos.
  - **Value 0 :** normal
  - **Value 1 :** présentant une anomalie de l'onde ST-T (inversions de l'onde T et/ou élévation ou dépression du segment ST de > 0,05 mV).
  - **Value 2 :** montrant une hypertrophie ventriculaire gauche probable ou certaine selon les critères d'Estes.
- **thalach :** fréquence cardiaque maximale atteinte.
- **target :** 0= moins de risque de crise cardiaque 1= plus de risque de crise cardiaque.

3.4.3 L'exploration et visualisation des données de heart

Du rapport HTML dans la figure ci-dessous, on va extraire toutes les informations de base sur notre dataset pour chercher les cases vides, les lignes dupliquées et les colonnes de type chaine de caractères (Figure 3.4) :

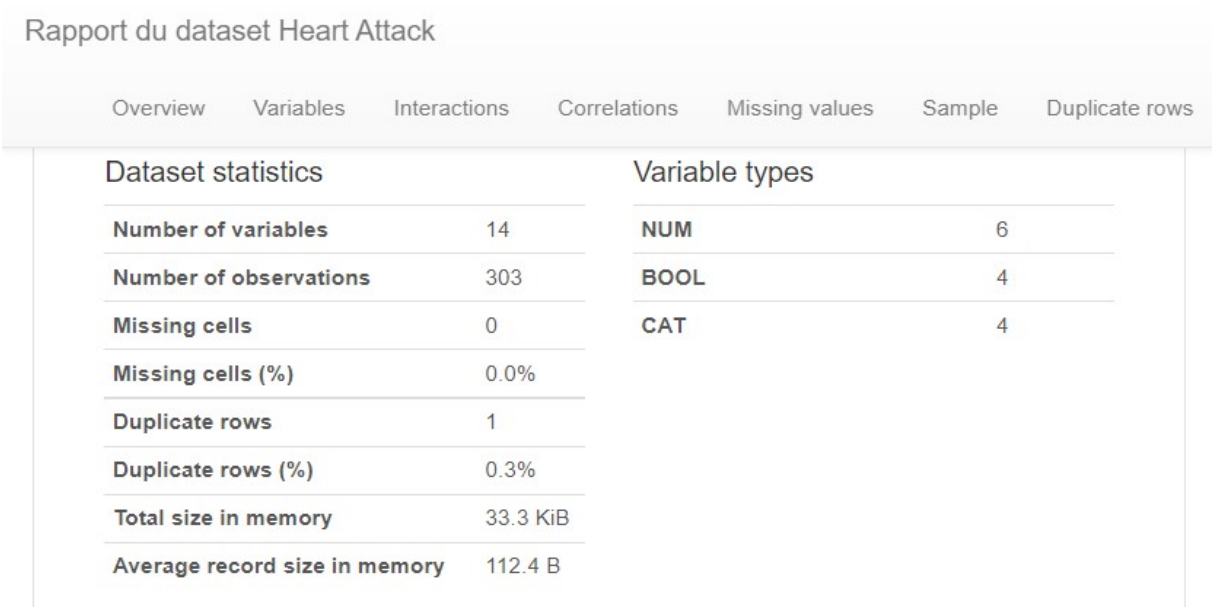

FIGURE 3.4 – Le rapport HTML du jeu de données heart.

Les informations extraites

— Le nombre des colonnes : 14 colonnes.

— Le nombre des observations : 303 observations.

— Le nombre et le type de variables :

  — 6 variables de type numérique.

  — 4 variables de type booléen.

  — 4 variables de type catégoriel.

— Les valeurs manquantes : 0.

— Le pourcentage (%) des valeurs null pour chaque variable : 0%.

— Lignes en double : une ligne.

— La taille de l'ensemble de données : 112.4 KB.

3.4.4 Nettoyage et pré-traitement du jeu de données heart

— Supression des lignes en double

La figure 3.5 suivante représente le dataset après la supression des lignes dupliquées.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

302 rows × 14 columns

FIGURE 3.5 – Dataset après la supression des lignes dupliquées.

3.5 Approche proposée

Notre methode consiste à ajouter des lignes à notre dataset en générant de fausses valeurs, ce qui nous permet de perturber l'ensemble de données original en ajoutant du bruit, ce qui nous donnera ce qu'on appelle un dataset perturbé.

Pour attiendre notre objectif, nous avons pensé à deviser notre jeu de données selon les attributs des colonnes cible (dans notre cas, la colonne cible a deux attributs 0 et 1, donc on aura deux jeu de données : le premièr a juste l'attribut 1 comme colonne cible, et le deuxième a juste l'attribut 0 comme colonne cible), puis nous déterminons la plus grande valeur et la plus petites valeur de chaque colonne dans les deux jeu de données résultant, puis on ajoute un nombre limité de valeurs dans les deux sous ensemble de données et cela d'une manière a toujours respecter l'ensemble de données original, entre les valeurs maximale et minimale des colonnes, sans pour autant changer les attributs de colonne cible.

L'idée de base qui nous permet de garder l'utilité des données dans le nouveau dataset est liée aux nombre de valeurs ajoutées, que nous difinissons par le quotient de la division du taille de jeu de données qui a l'attribut 1 sur la taille de jeu de données qui a l'attribut 0, ce quotient doit être le même pour les deux datasets (l'original et le nouveau dataset perturbé). Nous concluons notre méthode par la classification naive bayesienne GaussianNB pour pouvoir évaluer notre nouveau dataset perturbé.

La figure 3.6 suivante représente l'architecture de notre proposition :

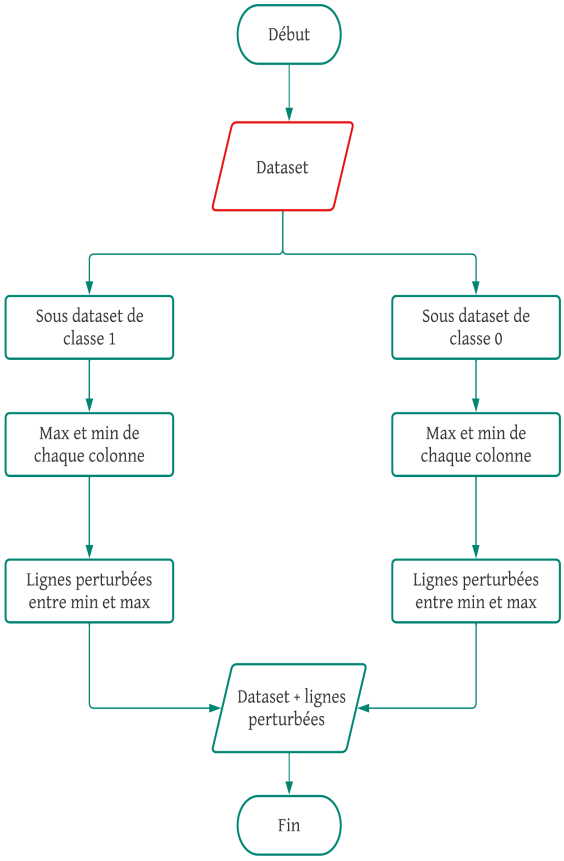

FIGURE 3.6 – Architecture de notre proposition.

3.6 Explication de l'approche

Etape 1 : Dataset

Cette étape nous permet de créer notre jeu de données, et cela a déjà été expliqué dans la section 3.4.

Etape 2 : Pré-traitement

Cette étape est déjà expliqué dans la section 3.4.

Etape 3 : l'approche d'anonymat proposée

1. **Diviser le jeux de données selon les classes**

   La figure 3.7 suivante représente la partie du dataset original qui a juste la classe 0.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
165	67	1	0	160	286	0	0	108	1	1.5	1	3	2	0
166	67	1	0	120	229	0	0	129	1	2.6	1	2	3	0
167	62	0	0	140	268	0	0	160	0	3.6	0	2	2	0
168	63	1	0	130	254	0	0	147	0	1.4	1	1	3	0
169	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

138 rows × 14 columns

FIGURE 3.7 – Partie du dataset original qui a juste la classe 0.

   La figure 3.8 suivante représente la partie du dataset original qui a juste la classe 1.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
159	56	1	1	130	221	0	0	163	0	0.0	2	0	3	1
160	56	1	1	120	240	0	1	169	0	0.0	0	0	2	1
161	55	0	1	132	342	0	1	166	0	1.2	2	0	2	1
162	41	1	1	120	157	0	1	182	0	0.0	2	0	2	1
163	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1

164 rows × 14 columns

FIGURE 3.8 – Partie du dataset original qui a juste la classe 1.

2. **Determiner le maximum, et le minimum des valeurs pour chaque colonne dans les deux jeux de données**

   La figure 3.9 représente le maximum et le minimum des valeur pour chaque colonne dans les deux jeux de données.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
max1	76	1	3	180	564	1	2	202	1	4.2	2	4	3	1
min1	29	0	0	94	126	0	0	96	0	0.0	0	0	0	1
max0	77	1	3	200	409	1	2	195	1	6.2	2	4	3	0
min0	35	0	0	100	131	0	0	71	0	0.0	0	0	0	0

FIGURE 3.9 – Le maximum et le minimum des valeur pour chaque colonne dans les deux jeux de données.

3. **Générer un nombre limité de lignes (d'une manière qui respecte l'ensemble de données original) entre le min et le max de chaqu'une des colonnes**

   La figure 3.10 suivante représente les lignes ajoutées.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	41	0	2	174	352	0	1	108	0	3.4	1	2	1	1
1	43	0	1	165	459	0	1	177	0	0.1	0	2	2	1
2	68	0	0	94	209	0	1	150	0	0.2	0	1	1	1
3	53	0	1	146	352	0	0	101	0	4.0	0	1	0	1
4	47	0	0	175	281	0	1	124	0	0.9	0	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
64	39	0	1	181	181	0	1	108	0	5.5	0	2	1	0
65	37	0	2	170	164	0	1	132	0	2.3	0	0	1	0
66	73	0	1	183	404	0	0	194	0	2.8	0	1	1	0
67	70	0	0	147	133	0	0	193	0	3.1	0	1	2	0
68	38	0	2	175	358	0	1	128	0	0.3	0	1	0	0

151 rows × 14 columns

FIGURE 3.10 – Les lignes ajoutées.

4. **Ajouter les lignes générées au jeu de données original**

   La figure 3.11 représente le dataset anonymisée.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	76	0	0	198	386	0	0	72	0	0.8	1	0	1	0
1	65	0	1	118	367	0	0	126	0	4.2	0	0	0	0
2	52	0	2	152	339	0	0	187	0	5.2	0	2	0	0
3	44	0	1	106	341	0	1	147	0	3.6	0	1	0	0
4	51	0	1	144	313	0	1	71	0	2.8	0	3	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

453 rows × 14 columns

FIGURE 3.11 – Dataset anonymisée.

Etape 4 : Appliquer la méthode de classification naive bayes (GaussianNB) sur le jeux de données perturbé :

Tout d'abord, on a séparé les colonnes en variables dépendantes et indépendantes (ou caractéristiques et étiquette). Ensuite, on a divisé ces variables en ensemble d'entraînement et de test.

- **Verticalement :** on a divisé Notre dataset de manière verticale en deux variables :
  - **La variable y :** Représente la colonne cible "Label" c'est ce qu'on souhaite que la machine apprenne à prédire .
  - **La variable X :** Représente les facteurs "Features" (X1 ,X2,...,X13), c'est toute les colonnes de dataset sauf la colonne cible output .
- **Horizontalement :** On a divisé Notre dataset de manière horizontale en deux parties :
  - **La partie train :** C'est la base d'entrainement "Training set" dont laquelle le modèle fait son apprentissage.
  - **La partie test :** C'est l'ensemble de test "Test set" dont laquelle nous testons notre modèle et évaluer sa performance.

Dans notre cas on a défini l'ensemble d'apprentissage par 70% (211 lignes) et l'ensemble de test par 30% (91 lignes) (voir figure 3.12).

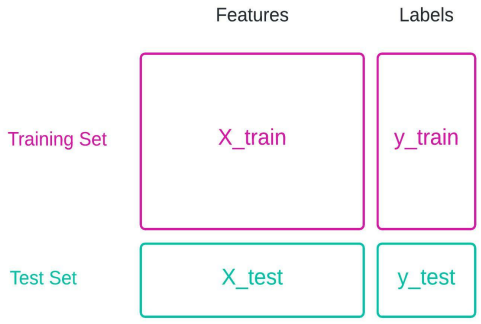

FIGURE 3.12 – La division de l'ensemble de données verticalement et horizontalement.

Maintenant, notre jeu de données est prêt pour appliquer la classification Gaussian naive bayes.

3.7 Environnements et outils de développement

Pour implémenter notre système, nous avons utilisé des environnements, des outils de langage et des bibliothèques :

3.7.1 Plate-formes et environnement

— **Anaconda :** Anaconda est une plate-forme gratuite et open source de science des données et d'apprentissage automatique pour les langages de programmation Python et R. L'idée de base d'Anaconda est de permettre aux personnes intéressées par ces domaines d'installer facilement tous (ou la plupart) des packages requis dans une seule installation [54][55].

— **Jupyter Notbook :** Il s'agit d'un environnement de développement interactif basé sur le Web qui permet aux utilisateurs de créer et de partager des symboles, des équations, des visualisations et des textes [56].

— **Google Colaboratory Notbook :**[1] Colaboratory, ou « Colab » en abrégé, est un produit de Google Research. Colab permet à quiconque d'écrire et d'exécuter

1. https ://colab.research.google.com/

du code Python arbitraire via le navigateur, et est particulièrement bien adapté à l'apprentissage automatique, à l'analyse de données et à l'éducation [57].

3.7.2 Langage de programmation

— **Python :** C'est un langage de programmation open source multiparadigme. Il libère les développeurs eux-mêmes des contraintes de formatage qui occupaient leur temps et permet aux développeurs de se concentrer sur ce qu'ils font au lieu de sur la façon de le faire. Ainsi, développer du code avec Python est plus rapide que d'autres langages [58].

3.7.3 Bibliothèques Utilisées

— **Panda :** Pandas est un package Python open source largement utilisé dans les tâches de science des données/d'analyse de données et d'apprentissage automatique. Pandas fonctionne bien avec de nombreux autres modules de science des données au sein de l'écosystème Python et est généralement inclus dans toutes les distributions [59].

— **Scikit-learn :** est une bibliothèque clé pour le langage de programmation Python qui est généralement utilisé dans les projets d'apprentissage automatique. Les concepts et fonctionnalités clés incluent [60] :
  — Algorithmes prenant en charge l'analyse prédictive allant de la simple régression linéaire à la reconnaissance de modèles de réseaux neuronaux.
  — Méthodes informatiques d'aide à la décision, notamment : classification, régression, clustering.
  — Interopérabilité avec les bibliothèques NumPy, pandas et matplotlib.

— **Numpy :** est une bibliothèque qui permet d'effectuer des calculs numeriques avec Python. Il s'agit d'une bibliothèque Python qui fournit un objet tableau multidimensionnel, divers objets dérivés (tels que des tableaux masqués et des matrices) et un assortiment de routines pour des opérations rapides sur des tableaux, y compris mathématiques, logiques, manipulation de forme, tri, sélection, transformées de Fourier discrètes, algèbre linéaire de base, opérations statistiques de base, simulation aléatoire et bien plus encore [61].

— **matplotlib :** est une bibliothèque complète pour créer des visualisations statiques,

animées et interactives en Python, elle génère des tracés, des histogrammes, des spectres de puissance, des diagrammes à barres, des diagrammes d'erreurs, des diagrammes de dispersion, etc., avec seulement quelques lignes de code [62].

— **Seaborn :** Seaborn est une bibliothèque qui offre la possibilité de résumer et de visualiser des données. Il permet de créer de jolis graphiques statistiques en Python. Cette bibliothèque fournit de nouvelles fonctionnalités qui améliorent l'exploration et la compréhension des données [63].

3.8 Resultats expérimentaux et évaluation

Dans cette partie , nous allons d'abord montrer la méthode d classification que nous allons l'utiliser pour classifier notre jeu de données, puis évaluer les performances des résultats de classification de notre méthode par rapport aux resultats de classification du jeu de données original, en utilisant divers mesures d'évaluation.

Méthode de classification utilisée

Pour classifier le nouveau jeu de données obtenu de l'approche proposée, on a choisi la méthode de classification Gaussian naïve Bayes (GNB).

Le GNB est un algorithme d'apprentissage supervisé qui utilise le théorème de Bayes comme cadre pour classer les observations dans l'une des classes prédéfinies en fonction des informations fournies par les variables prédictives. Les classificateurs GNB estiment les probabilités conditionnelles qu'une observation appartienne à une classe particulière étant donné les valeurs des variables prédictives sous l'hypothèse que les variables prédictives sont conditionnellement indépendantes de la classe, et donc (naïvement) ne prennent pas en compte la covariance entre les variables prédictives [64].

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

Ou le μ est la moyenne définis par :

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.2}$$

Et le σ est la variance, qui est donnée par l'expresssion suivante :

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5} \tag{3.3}$$

Les avantages du classificateur GNB

- Un modèle rapide et flexible donne des résultats très fiables et fonctionne bien avec des données volumineuses.
- Il n'y a pas besoin de passer beaucoup de temps pour la formation.
- Fournit une meilleure performance de garde en éliminant les spécifications insignifiantes [48].

3.8.1 Matrice de confusion

La matrice de confusion est un outil permettant de résumer les performances d'un algorithme de classification. Il nous donne un résumé des prédictions correctes et incorrectes ventilées pour chaque catégorie. Le résumé est représenté sous forme de tableau.

Quatre types de résultats sont possibles lors de l'évaluation des performances d'un modèle de classification. Ces quatre résultats sont donnés ci-dessous :

- **TP (True positives) :** Les vrais positifs se produisent lorsque nous nous attendons à ce qu'une observation appartienne à une certaine catégorie et que l'observation appartienne réellement à cette catégorie.
- **TN (True negatives) :** Les vrais négativités se produisent lorsque nous nous attendons à ce qu'une observation n'appartienne pas à une certaine catégorie et que l'observation appartienne en fait à cette catégorie.
- **FP (False positives) :** Les faux positifs se produisent lorsque nous nous attendons à ce qu'une observation appartienne à une certaine catégorie, mais que l'observation n'appartient pas réellement à cette catégorie.
- **FN (False negatives) :** Les faux négatifs se produisent lorsque nous nous attendons à ce qu'une observation n'appartienne pas à une catégorie particulière, mais que l'observation appartient en fait à cette catégorie.

Les figures [3.13] [3.14] suivantes montrent les matrices de confusion de chaque ensemble de données (original et anonymisé) :

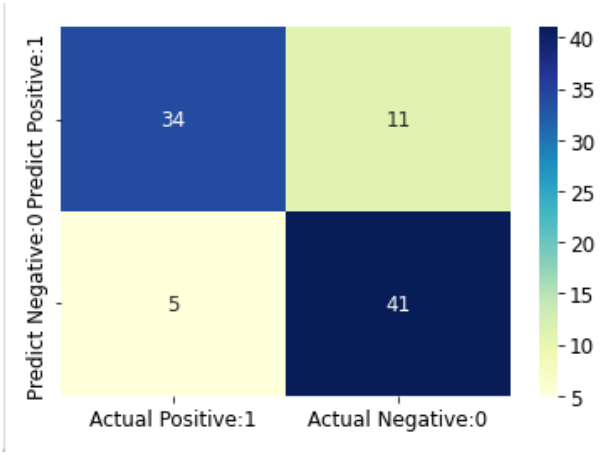

FIGURE 3.13 – Matrice de confusion du jeu de données original.

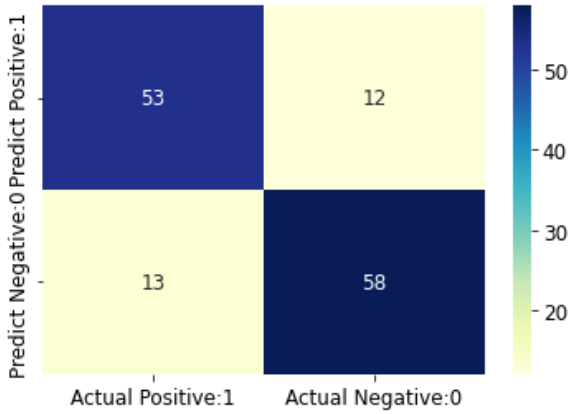

FIGURE 3.14 – Matrice de confusion du jeu de données anonymisé.

Afin que nous pouvons comparer les matrices de confusion des deux jeux de données, nous avons suggéré de générer le pourcentages des valeurs (FP, FN, TN, TP) dans les deux matrices par rapport à la base de test de chaque dataset(voir Figures 3.15 et 3.16).

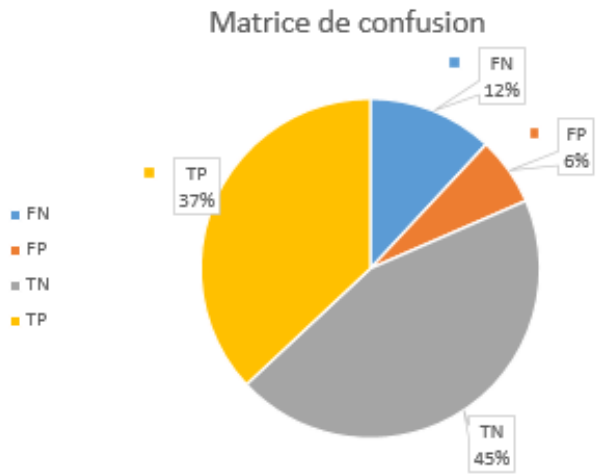

FIGURE 3.15 – Secteur des proportions des valeurs de matrice de confusion (original).

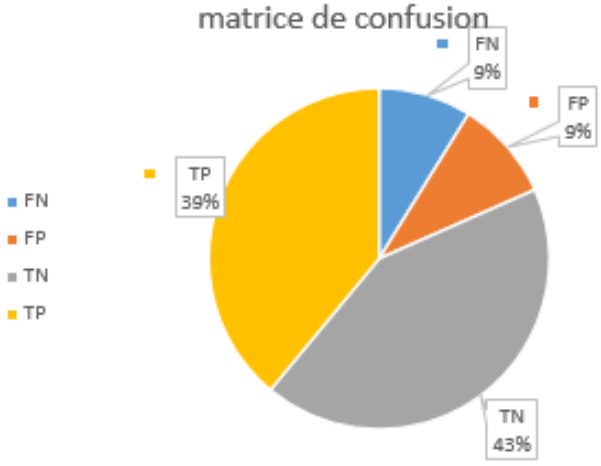

FIGURE 3.16 – Secteur des proportions des valeurs de matrice de confusion (anonymisé).

D'après les résultats présentés dans les figures ci-dessus, nous remarquons que les pourcentages de valeurs sont presque similaires dans les deux jeux de données.

Les figures [3.17][3.18] suivantes représentent les rapports de classification des mesures d'évaluation que nous allons mentionnées précédemment pour le jeu de données original et anonymisé :

	precision	recall	f1-score	support
0	0.87	0.76	0.81	45
1	0.79	0.89	0.84	46
accuracy			0.82	91
macro avg	0.83	0.82	0.82	91
weighted avg	0.83	0.82	0.82	91

```
[[34 11]
 [ 5 41]]
```

FIGURE 3.17 – Les résultats des attributs d'évaluation du dataset original.

	precision	recall	f1-score	support
0	0.80	0.82	0.81	65
1	0.83	0.82	0.82	71
accuracy			0.82	136
macro avg	0.82	0.82	0.82	136
weighted avg	0.82	0.82	0.82	136

```
[[53 12]
 [13 58]]
```

FIGURE 3.18 – Les résultats des attributs d'évaluation du dataset anonymisé.

A travers les figures ci-dessus, nous constatons que les résultats de mesures d'évaluation des deux dataset sont presque égaux.

3.8.2 Accuracy

Définition : C'est le rapport du nombre de prédictions correctes au nombre total d'échantillons d'entrée. Cela fonctionne bien s'il y a un nombre égal d'échantillons appartenant à chaque classe [65].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

Resultat

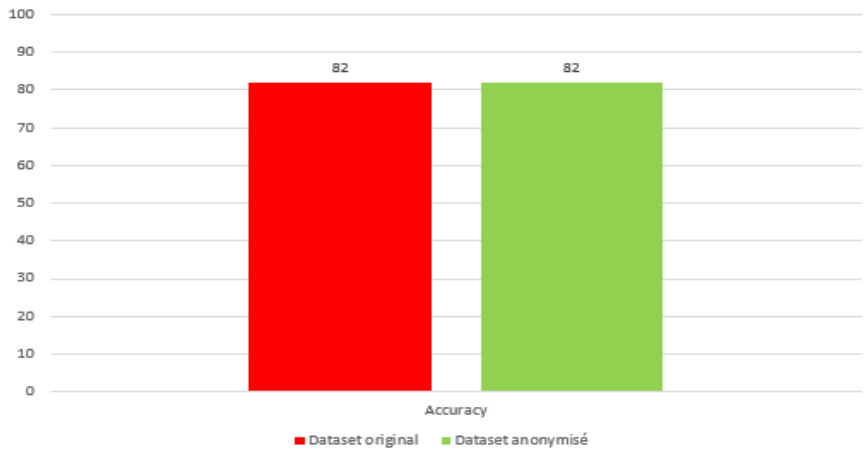

FIGURE 3.19 – Comparaison d'accuracy entre le dataset original et anonymisé.

L'accuracy du dataset original est de 82% et l'accuracy du dataset anonymisé est de 82%, donc nous constatons que ces résultats sont similaires.

3.8.3 Précision

Définition : la précision fait référence au rapport entre les vraies prédictions positives et le nombre total de prédictions positives [65].

$$Precision = \frac{TP}{TP + FP} \tag{3.5}$$

Resultat

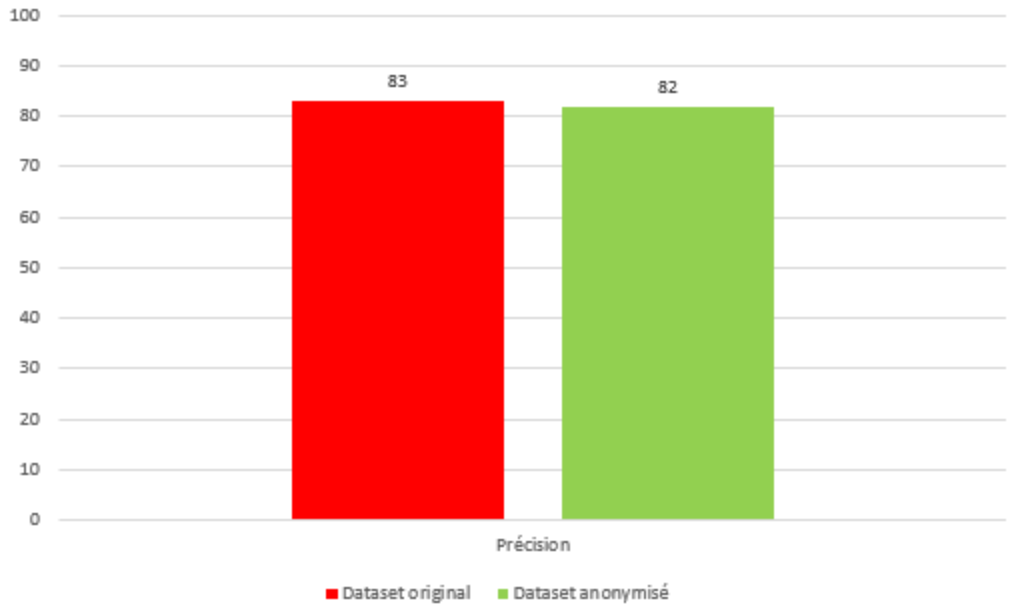

FIGURE 3.20 – Comparaison de précision entre le dataset original et anonymisé.

D'après la figure 3.20 précédentes, nous constatons que la précision du dataset original est de 83% et la précision du dataset anonymisé est de 82%, alors on peut dire que ces résultats sont très proches.

3.8.4 Rappel(Recall)

Définition : le rappel est un paramètre qui mesure le nombre de prédictions positives valides sur le nombre total de données positives [65].

$$Recall = \frac{TP}{TP + FN} \tag{3.6}$$

Resultat

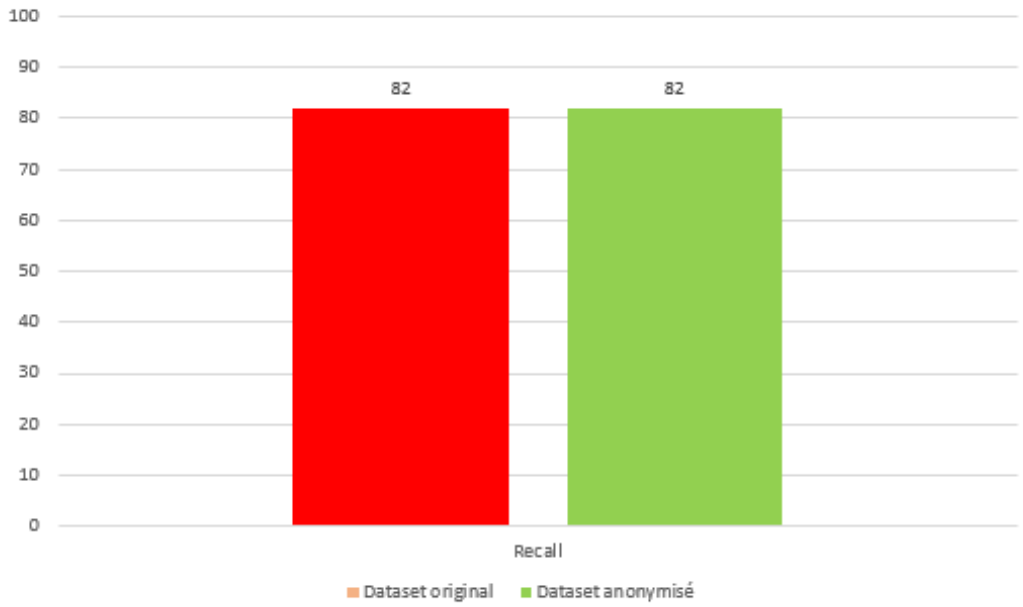

FIGURE 3.21 – Comparaison de rappel entre le dataset original et anonymisé.

On constate, à travers la figure 3.21 ci-dessus, que le rappel de jeu de données original est de 82% et pour l'anonymisé est de 82%, donc ils sont égaux.

3.8.5 F1 score

Définition : le score F1 est un moyen harmonique de précision et de rappel. C'est le double du produit de ces deux paramètres par leur somme [65].

$$F1 = 2 * \frac{Precision * recall}{Precision + recall} \tag{3.7}$$

Resultat

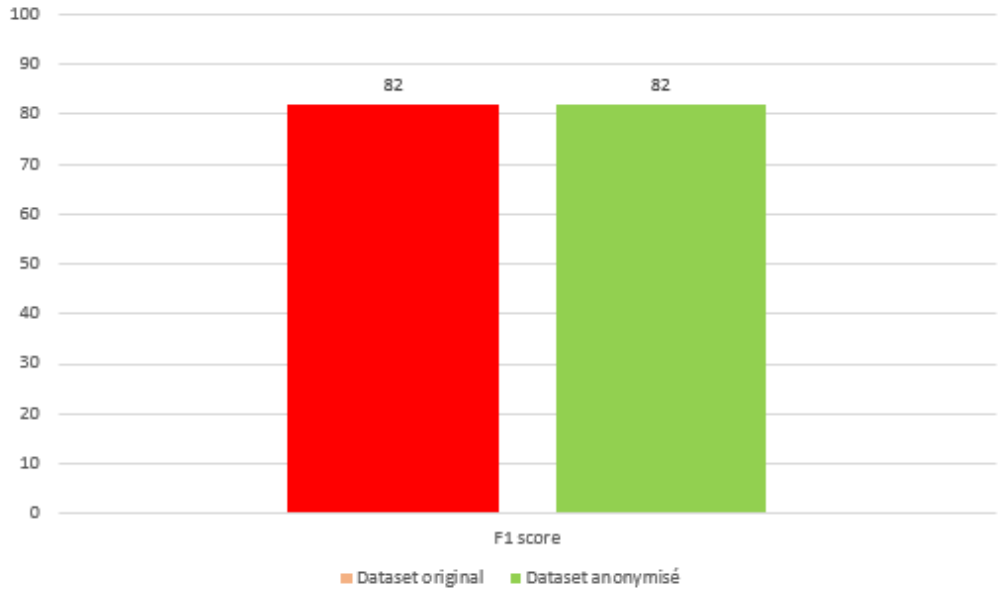

FIGURE 3.22 – Comparaison de F1 score entre le dataset original et anonymisé.

A travers la figure 3.22 ci-dessus, nous remarquons que les deux ensembles de données sont egaux.

3.8.6 Entropie

Définition : Est une fonction mathématique qui, intuitivement, correspond à la quantité d'information contenue ou délivrée par une source d'information.

$$E(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-) \tag{3.8}$$

— S : Ensemble de données (Heart Attack).

— p+ : la proportion de classe 1.

— p- : la proportion de classe 0.

Resultat

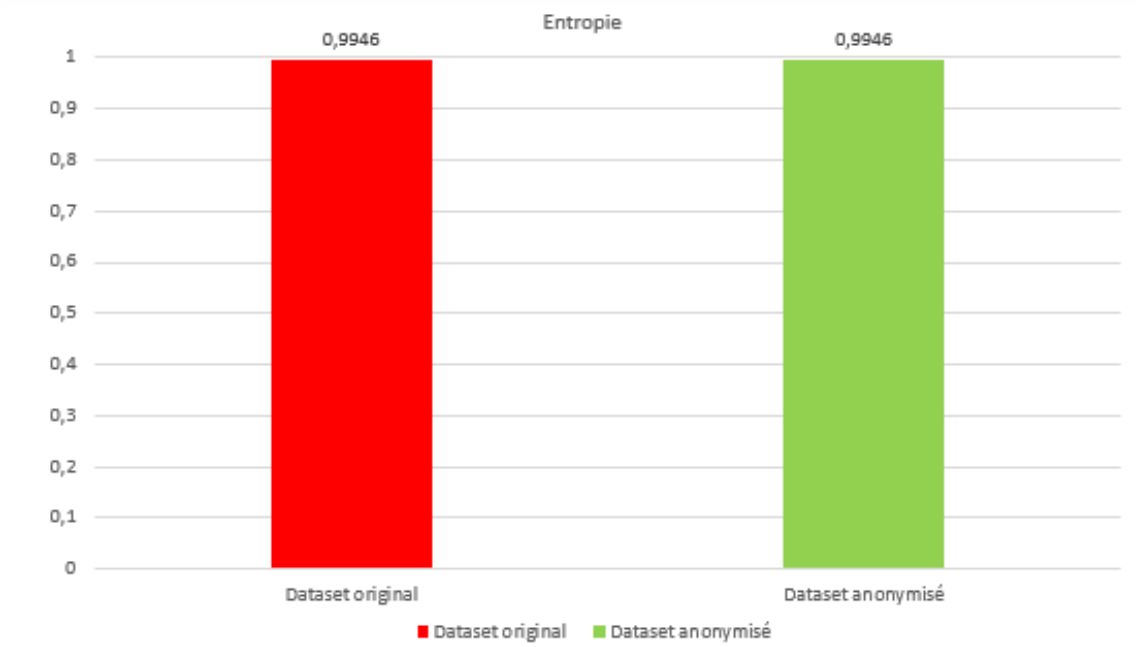

FIGURE 3.23 – Comparaison d'entropie entre le dataset original et anonymisé.

D'après la figure 3.23 précédente, on peut remarquer que l'entrepie égal à 0,9946 pour les deux ensembles de données.

3.8.7 Courbe ROC

Une courbe ROC (Receiver Operating Characteristic Curve) est un graphique montrant les performances d'un modèle de classification à tous les seuils de classification. Elle trace la sensitivité par rapport au spécificité à différents seuils de classification. L'abaissement du seuil de classification classifie plus d'éléments comme positifs, augmentant ainsi à la fois les faux positifs et les vrais positifs [66]. Les figures suivantes montrent les courbes ROC de dataset original et anonymisé :

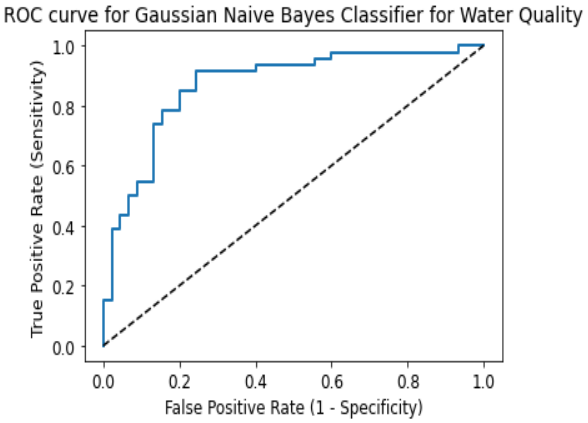

FIGURE 3.24 – Courbe ROC du dataset original.

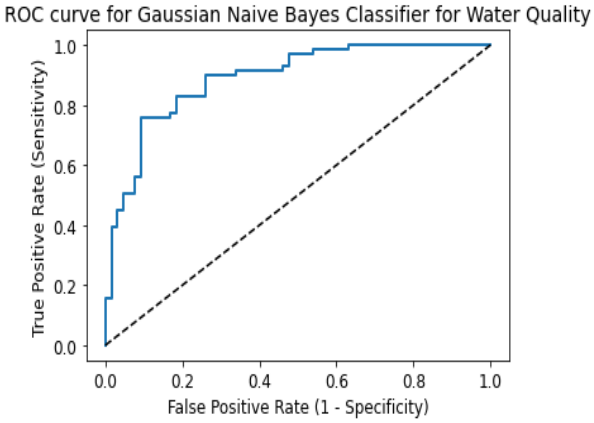

FIGURE 3.25 – Courbe ROC du dataset anonymisé.

D'après les figures [3.24] [3.25] ci-dessus, nous remraquons que les courbe ROC des deux jeux de données sont presque identiques.

3.8.8 Autres évaluation

— Les figures [3.26] [3.27] ci-dessous montrent que le nombre de lignes ajoutées garde le même pourcentage des classes par rapport au jeu de données original :

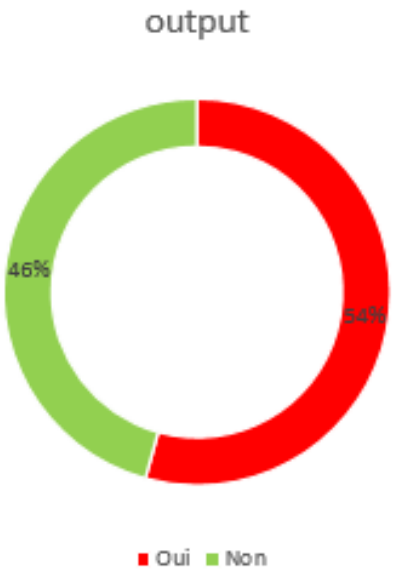

FIGURE 3.26 – Proportion des attributs de colonne cible (original).

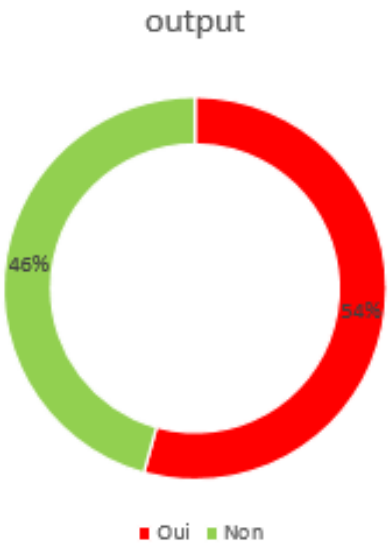

FIGURE 3.27 – Proportion des attributs de colonne cible (anonymisé).

D'après les figures, nous constatons que les pourcentages des attributs du colonne

cibles sont égaux dans les deux datasets.

— Les figures suivantes montrent des histogrammes des probabilités prédites (0,1) pour les jeux de données originals et anonymisés :

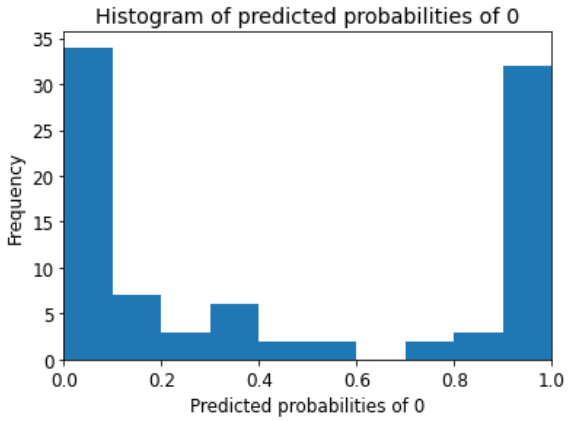

FIGURE 3.28 – Histogramme pour les probabilités prédites de 0 (original).

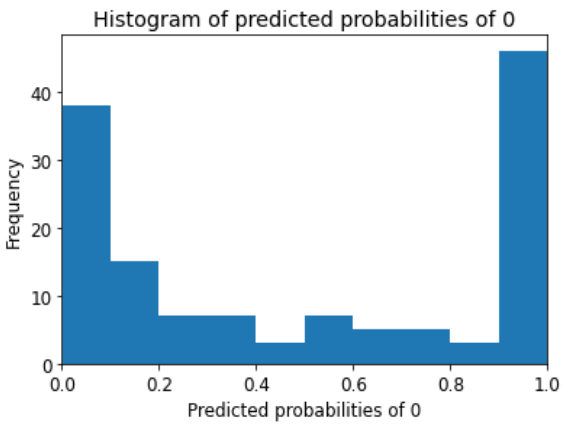

FIGURE 3.29 – Histogramme pour les probabilités prédites de 0 (anonymisé).

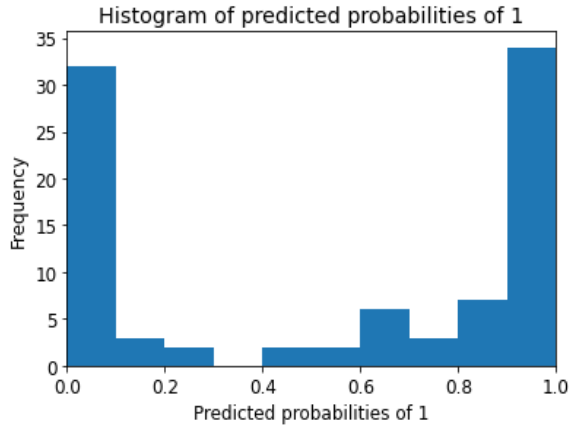

FIGURE 3.30 – Histogramme pour les probabilités prédites de 1 (original).

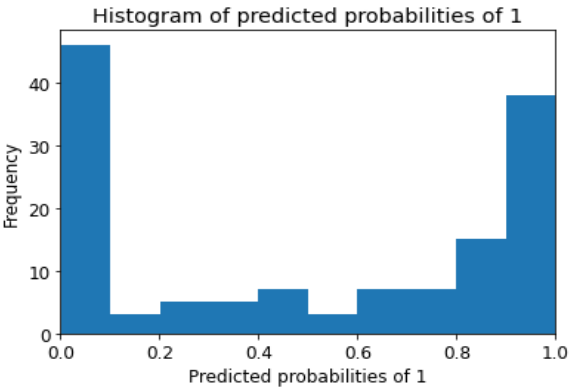

FIGURE 3.31 – Histogramme pour les probabilités prédites de 1 (anonymisé).

D'après les histogrammes montrés ci-dessus, nous remarquons que les histogrammes sont presque identiques pour les deux cas, pour les deux jeux de données.

— Les figures suivantes montrent des prédiction de lignes à l'aide de l'ensemble de données original et l'ensemble anonymisé :

```
X_newval=np.array([60,0,1,130,237,0,0,165,1,1.2,1,1,3])
y_pred=NBModel.predict([X_newval])
y_pred
array([0], dtype=int64)
```

FIGURE 3.32 – Prédiction avec le dataset original.

```
X_newval=np.array([60,0,1,130,237,0,0,165,1,1.2,1,1,3])
y_pred_anony=NBModel.predict([X_newval])
y_pred_anony
array([0], dtype=int64)
```

FIGURE 3.33 – Prédiction avec le dataset anonymisé.

```
X_newval=np.array([60,0,1,141,227,1,0,175,1,2.3,1,1,2])
y_pred=NBModel.predict([X_newval])
y_pred
array([0], dtype=int64)
```

FIGURE 3.34 – Prédiction avec le dataset original.

```
X_newval=np.array([60,0,1,141,227,1,0,175,1,2.3,1,1,2])
y_pred_anony=NBModel.predict([X_newval])
y_pred_anony
array([0], dtype=int64)
```

FIGURE 3.35 – Prédiction avec le dataset anonymisé.

Nous pouvons constater, à travers les figures précedentes, que les prédictions sont les mêmes dans les deux jeux de données, pour les deux lignes prédictées.

3.9 Conclusion

Dans cette partie, nous avons presenté la méthode utilisée pour accomplir notre travail en detaillant ses differentes étapes. Tout d'abord, on a montioné le jeu de données et les étapes de preparation des donnees afin de s'adapter a l'objectif de notre étude. Ensuite nous avons expliqué notre approche de maniére précise et détaillée afin de justifier la solution que nous avons proposée. puis nous avons recensé les differents environements et outils de développement utilisé pour l'implémentation de notre approche et son évaluation.

D'apres les figures précédentes nous constatons que les résultats de classification de l'ensemble de données anonymisé étaient très proches des résultats de l'ensemble de données original, donc on peut conclure que l'approche proposée maintient l'exactitude et l'utilité des informations, ainsi qu'elle est facile, simple et efficace contre les attaques et la ré-identification.

Conclusion générale et perspectives

Durant le travail de notre mémoire, nous avons présenté les différentes étapes de la réalisation de notre approche pour la classification préservant la vie privé en utilisant les méthodes d'anonymat et machine learning.

Dans un premier temps on a présenté des généralités sur la préservation de la confidentialité et l'anonymat pour les données publiées. Au départ, nous avons commencé par mentionner les définitions importantes liées à l'approche de l'anonymat et de la vie privée, les différents types, puis la classification et ses techniques et divers travaux liés à notre domaine de recherche.

Notre principal objectif était de proposer une méthode qui permet d'anonymiser une dataset contre les risques de la ré-identification, tout en respectant la précision et la qualité des données, alors nous avons proposé de perturber le jeu de données en ajoutant des lignes d'une façon a respecter les valeurs du dataset original, tout en respectant les propotions des différents individus (positifs et négatifs) du dataset original pour garantir que l'utilité des données ne sera pas perdue. Nous avons ensuite évalué la pertinence de notre nouveau dataset avec l'original en appliquant la classification Naive Bayes et on a comparé les deux modèles résultants des deux dataset.

Après toutes les expèriences que nous avons faite durant ce travail, nous constatons que les résultats de classification de notre proposition et de jeu de données original sont presque les mêmes, donc nous pouvons dire que la qualité des données anonymisées est sans perte significative de l'utilité par rapport aux données originale, ce qui montre l'efficacité de notre proposition.

De plus, avant de passer aux perspectives, ce projet a été utile à plusieurs niveaux :

— Sur le plan technique, nous avons eu l'opportunité d'enrichir nos connaissances des outils et environnements de développement tels qu'Anaconda, Jupyter Notbook et Google Colaboratory. Ainsi, on a eu l'opportunité de maitriser le langage de programmation Python.

— Sur le plan personnel, cette expérience pratique nous a permis de découvrir le milieu professionnel avec tout ce qu'il exige de discipline et de responsabilité ce qui va nous aider à enrichir notre carrière professionnelle dans des meilleurs niveaux d'appartenance et savoir.

Perspectives

Notre travail est certe loin d'être complet et parfait, c'est pourquoi nous envisageons pour les travaux futures les perspectives suivantes :

— Valider notre approche en testant des attaques de confidentialité.

— Développer notre méthode pour un ensemble de données contenant plus de deux attributs dans la classe cible.

— Proposer des solutions de préservation de la vie privée à d'autres algorithmes de classification.

Bibliographie

[1] Benjamin Nguyen. Techniques d'anonymisation. Statistique et Société, Société française de statistique, 2014, 2 (4), pp.53-60.

[2] Benjamin Nguyen, 29/04/2021, courses protegez les donnees personnelles/manipulez les modeles d'anonymisation.

[3] https ://towardsdatascience.com/, consulté le 27/02/2022.

[4] https ://eric.univ-lyon2.fr/ ricco/cours/slides/svm.pdf, consulté le 02/03/2022.

[5] David Roman, Jun 3, 2017, Réseaux neuronaux pour les nuls.

[6] https ://fleid.net/category/techno/ssrs/, consulté le 21/03/2022.

[7] Beynon-Davies, P. (2004). Distributed data. In : Database systems. Palgrave, London.

[8] T.Boudheb, these de doctorat, 2019, Privacy Preserving Classification of Biomedical Data.

[9] M.Sahu, D.Gountia, N.Samal, april 2013, privacy preservation decision tree based on data set complementation.

[10] https ://projeduc.github.io/intro_apprentissage_automatique/, consulté le 25/02/2022.

[11] A.Aid, 2021, cours 6 module Fouille de données, Master1ISIL université de bouira.

[12] https ://www.purevpn.com/blog/what-is-internet-privacy-scty/, consulté le 01/02/2022.

[13] Artrit.Ajdini, 30 septembre 2019, Étude et conception d'un service assurant l'anonymat, Genève.

[14] https ://fr.differbetween.com/article/difference-between-confidentiality-and-anonymity, consulté le 02/02/2022.

[15] CJUE, 6 octobre 2020, c-623/17 privacy international.

[16] Charlotte GALICHET, 25 mai 2020, la CNIL fait le point sur les techniques d'anonymisation des données personnelles.

[17] L. Sweeney, Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, International Journal on Uncertainty, Vuzziness, and Knowledge-based Systems, Vol. 10, No. 5 (2002), pp. 571–588.

[18] L. Xu, C. Viang, J. Wang, J. Yuan and Y. Ren, "Information Security in big data : Privacy and data mining," in IEEE access, vol. 2, pp. 1149-1176, 2014, doi : 10.1109/ACCESS.2014.2362522.

[19] A. Majeed and S. Lee, "Anonymization Techniques for Privacy Preserving Data Publishing : A Comprehensive Survey," in IEEE Access, vol. 9, pp. 8512-8545, 2021, doi : 10.1109/access.2020.3045700.

[20] https ://www.privitar.com/glossary/quasi-identifier/, consulté le 03/06/2022.

[21] https ://www.datanaos.com/techniques_anonymisation, consulté le 05/02/2022.

[22] Techniques d'anonymisation,Benjamin NGUYEN Insa1 Centre Val de Loire et Inria2 Paris-Rocquencour/ss.pdf.

[23] C. Dwork, 9-16 2006, "differential privacy," in 33rd international colloquium on automata, languages and programming (icalp 2006), venice, italy,jul.

[24] Vaidya, J., Shafiq, B., Basu, A., Hong, Y, 2013, Differentially private naive bayes classification.

[25] Mohammad Waseem, How To Implement Classification In Machine Learning ? , Jan 05,2022.

[26] T.N.Phyu, 2009, Survey of Classification Techniques in data mining.

[27] Abbas Rizvi, 2010, ID3 Algorithm, CS157 B, Spring.

[28] J.Vaidya, C.W. Clifton, Y.M.Zhu, 2006, privacy preserving data mining.

[29] Vaidya, J., Kantarcıoğlu, M., Clifton, C, 2007, Privacy-preserving naïve bayes classification.

[30] https ://www.lovelyanalytics.com/2018/10/04/classification-bayesienne-naive-comment-ca-marche/, consulté le 27/02/2022.

[31] G.Kesavaraj, Dr.S.Sukumaran, 2013, a study on classification techniques in data mining.

[32] K.Harifi, 21 sept. 2019, Bien comprendre l'algorithme des k-plus proches voisins (fonctionnement et implémentation sur r et python).

[33] R. Sadiq, M.J. Rodriguez, 2011, empirical models to predict disinfection by-products (dbps) in drinking water encyclopedia of environmental health.

[34] Gutwirth, S., Hildebrandt, M. (2008). Profiling the European citizen : Cross-disciplinary perspectives. Springer.

[35] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren, 2014, information security in big data : Privacy and data mining.

[36] J.Dowd, S.Xu, W.Zhang, 2006, privacy-preserving decision tree mining based on random substitutions.

[37] Brickell, J, Shmatikov, V. (2009). privacy-preserving classifier learning. lecture notes in computer science.

[38] P. K. Fong and J. H. Weber-Jahnke, march 2012, privacy preserving decision tree learning using unrealized data sets.

[39] Sheela, M. A., & Vijayalakshmi, K, 2013, a novel privacy preserving decision tree induction.

[40] C. CLIFTON, M. KANTARCIOGLOU, X. LIN, Y. ZHU, 2002, "Tools for privacy pre- serving distributed data mining".

[41] H. LI ,D. JI, D. FENG, B. LI, July 2004, "Oblivious polynomial evaluation".

[42] U.Sangwan, REVIEW ON PRIVACY ISSUES IN BIG DATA USING DATA MINING TECHNIQUES.

[43] M.Skarkala, M.Maragoudakis, s.gritzalis, l.mitrou, 2011, privacy preserving tree augmented naïve bayesian multi–party implementation on horizontally partitioned databases.

[44] Shi Hong-bo, Wang Zhi-Hai, Huang Hou-Kuan, Jing Li-Ping. (n.d.)., 2002, text classification based on the tan model.

[45] T.BENKHELIF, novembre 2018, THESE DE DOCTORAT DE" publication de données individuelles respectueuse de la vie privée", UNIVERSITE DE NANTES.

[46] J. Vaidya, H. Yu, and X. Jiang, 2008, 'privacy-preserving svm classification'.

[47] https ://www.geeksforgeeks.org/major-kernel-functions-in-support-vector-machine-svm/, consulté le 28/03/2022.

[48] Akkaya, Berke Çolakoğlu, Nurdan. (2019). Comparison of Multi-class Classification Algorithms on early diagnosis of heart diseases.

[49] Vaidya, J., Yu, H., Jiang, X. (2007). Privacy-preserving SVM classification. knowledge and information systems,.

[50] https ://sites.utexas.edu/annapanyupeng/2015/07/14/gaussian-kernel/, consulté le 29/03/2022.

[51] Lin, K.-P., & Chen, M.-S, 2011, On the design and analysis of the privacy-preserving svm classifier.

[52] https ://sites.utexas.edu/annapanyupeng/2015/07/14/gaussian-kernel/, consulté le 07/06/2022.

[53] https ://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset, consulté le 06/06/2022.

[54] https ://linuxhint.com/anaconda-python-tutorial/, consultté le 01/06/2022.

[55] https ://www.venturelessons.com/what-is-anaconda/, consulté le 01/06/2022.

[56] https ://geekflare.com/fr/jupyter-notebook-basics, consulté le 01/06/2022.

[57] https ://research.google.com/, consulté le 06/06/2022.

[58] https ://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/, consulté le 01/06/2022.

[59] https ://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/, consulté le 01/06/2022.

[60] https ://www.activestate.com/resources/quick-reads/what-is-scikit-learn-in-python/, consulté le 01/06/2022.

[61] https ://numpy.org/doc/stable/user/whatisnumpy.html, consulté le 01/06/2022.

[62] https ://datascientest.com/matplotlib, consulté le 01/06/2022.

[63] https ://www.jedha.co/formation-python/librairie-seaborn, consulté le 01/06/2022.

[64] Griffis, J. C., Allendorfer, J. B., Szaflarski, J. P. (2016). voxel-based gaussian naïve bayes classification of ischemic stroke lesions in individual t1-weighted mri scans. journal of neuroscience methods, 257, 97–108. doi :10.1016/j.jneumeth.2015.09.019 10.1016/j.jneumeth.2015.09.019.

[65] https ://www.jedha.co/formation-ia/matrice-confusion, consulté 02/06/2022.

[66] https ://developers.google.com/machine-learning/crash-course/classification/roc-and-auc ?hl=fr, consulté le 06/06/2022.

[67] https ://www.netinbag.com/fr/internet/what-is-a-data-mining-classification.html, consulté le 27/02/2022.

[68] Waleed khalid Shihab, Sivaram Prasad, january 2022, "ARTIFICIAL NEURAL NETWORK" .

[69] Preeti narayan Baser, Jatinder Kumar, R. Saini,October 2012, Comparative Analysis of Issues Related to Centralized and Distributed Warehouse Architectures.

[70] https ://medium.com/@cdsjatin/tell-me-again-how-much-gram-is-in-the-matrix, consulté le 28/03/2022.