UNIVERSITÉ DU QUÉBEC À MONTRÉAL

TRANSPARENCE ET RESPONSABILITÉ DES ALGORITHMES DE PERSONNALISATION

RAPPORT DE RECHERCHE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE

PAR

ROSIN CLAUDE NGUEVEU

NOVEMBRE 2018

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

La grande masse de données disponibles a permis l'essor de techniques et algorithmes d'apprentissage automatique, de fouille et d'analyse de données. Ces techniques sont de nos jours très utilisées pour faciliter les prises de décisions et les recommandations, à tel point qu'elles sont devenues omniprésentes. Par exemple, dans le domaine juridique, les algorithmes sont utilisés pour prédire le taux de récidive d'un détenu (*Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)*, *Public Safety Assessment (PSA)*, etc.) [Center, 2016], ou encore les services en ligne tels que Amazon, Google, utilisent ces techniques afin de proposer des résultats pertinents aux usagers de leurs services. Les systèmes financiers développent et utilisent de telles techniques avec leurs transactions pour prédire le taux de risque associé aux individus [Mahmoud *et al.*, 2008]. Les entreprises quant à elles, pré-filtrent leur choix de candidats automatiquement [Faliagka *et al.*, 2012]. Plus récemment encore, dans le domaine de la sécurité ces algorithmes sont utilisés pour identifier et authentifier des personnes [Fathy *et al.*, 2015]. Les données utilisées par ces algorithmes varient en fonction de la finalité de l'application développée. Cependant, toutes ces techniques ont un point commun, elles nécessitent une connaissance très précise des individus [Carrascosa *et al.*, 2015] afin d'obtenir de bon taux de succès. Afin de subvenir à cette "nécessite des données personnelles", les entreprises collectent toutes les informations disponibles sur les individus, y compris celles qui n'auraient pas de lien direct avec leur application finale, dans l'anticipation des besoins ou des études futurs.

Une des techniques de collecte d'information est le profilage furtif des internautes mis en place par certains fournisseurs de contenus en ligne. [Englehardt et Narayanan, 2016] à l'aide de leur plateforme OpenWPM ont pu mettre en évidence l'existence de nouvelles

techniques de profilage, dont notamment la prise d'empreintes du spectre du matériel audio en jouant des sons inaudibles pour l'être humain. Dans la même optique, [Laperdrix *et al.*, 2016] ont fait appel au *crowdsourcing* afin de démontrer l'unicité des traces laissées par les outils navigations Internet : les configurations personnelles, les extensions installées, les historiques de navigations apportent peu de valeurs prises séparément, mais identifient de manière quasi unique les internautes lorsqu'elles sont misent en commun. Cette précision dans la collecte des informations soulève d'importantes questions liées au respect de la vie privée [Wallach, 2014], car le but de la collecte est rarement mentionné explicitement et le consentement éclairé des utilisateurs est rarement demandé.

En plus des enjeux liés à la collecte d'information, l'utilisation ubiquitaire de ces algorithmes pose des problèmes liés à la transparence et à la responsabilité. En effet, certains des algorithmes les plus performants et utilisés fonctionnent comme des boites noires qui «produisent des résultats en fonction des entrées» [Yu *et al.*, 2007]. La machine crée ainsi en interne une représentation des données lui permettant de répondre efficacement aux questions qui lui sont posées. Cependant, ce niveau d'abstraction reste incompréhensible pour l'être humain. [Ribeiro *et al.*, 2016] ont pu montrer qu'un algorithme peut être précis dans ses décisions en s'appuyant sur des éléments hors du cadre de réponses possibles pour ce contexte (différencier des chiens des huskies en s'appuyant sur des éléments du paysage), soulevant ainsi des questions de confiance dans les décisions du système et au système au système lui-même. D'autres exemples plus critiques, tels que dans le domaine de la santé où les décisions prises ont des conséquences importantes sur la vie des individus, montrent la nécessite d'avoir des moyens et méthodes permettant de rendre compréhensible la décision de l'algorithme [Dave Gershgorn, 2017 ; Davide Castelvecchi, 2016]. De plus, les nouvelles régulations telles que le Règlement Général sur Protection des Données personnelles (RGPD [1]) en Europe amplifient le besoin de

1. `https://gdpr-info.eu/`

transparence en offrant un panel de ressources à toute personne sujette au règlement, notamment par les articles 15 - le droit d'accès à la donnée, 17 - le droit à l'oubli et 18 - le droit à la restriction de traitement.

La transparence n'est pas le seul enjeu lié à l'omniprésence des outils de décision automatique. Des erreurs sont commises ou encore des problèmes sociaux de discriminations se reflètent dans ces outils automatiques. En reprenant l'exemple de la justice prédictive, une étude réalisée par ProPublica a mis en évidence le fait que l'outil COMPAS - utilisé dans le comté de Broward, est fortement biaisé envers les prévenus afro-américains [Angwin *et al.*, 2016]. Leur analyse a notamment mis en évidence la différence significative entre les taux de faux positifs, +91.26% chez les prévenus de couleurs que dans le reste de la population. Plusieurs autres exemples de problèmes liés au traitement discriminatoire envers certaines minorités ([Guynn, 2015 ; AATIF, 2017 ; Ingold et Soper, 2016]) existent dans la littérature. Ces problèmes peuvent être dus à des intentions malveillantes qui sont introduites dans le classifieur, à des erreurs humaines, ou généralement parlant, du fait que les données sur lesquelles l'algorithme est entraîné sont déjà biaisées. Le classifieur apprend et reproduit ainsi le biais dans les données d'entraînement. Par exemple, aux États-Unis, Amazon a récemment été critiqué à cause de son amélioration automatique du système de «livraison le même jour», qui a majoritairement exclu les zones à prédominance afro-américaine [Ingold et Soper, 2016]. Cette «erreur» est potentiellement causée par le fait que ces zones sont les moins riches et les résidents utilisent beaucoup moins le service Amazon. Ces zones sont ainsi très peu représentées dans l'ensemble de données dans le jeu de données, et l'algorithme qui est entraîné sur ces données n'a pas assez de connaissances pour prendre une décision. Ces exemples mettent en évidence l'insuffisance de données d'entraînement concernant certaines populations, conduisant ainsi le classifieur à ne se représenter que la population majoritairement présente.

Une solution potentielle serait donc d'augmenter la diversité des données utilisées pour

l'entraînement.

Une seconde solution potentielle serait de retirer les attributs jugés discriminants dans les données d'entraînement et une dernière serait de construire différents modèles, un pour chaque groupe considéré. Cependant, aucune de ces approches ne serait suffisante. La première est limitée par la difficulté d'obtenir des données variées, la seconde est confrontée au fait que certains attributs tout en apportant leur information propre, peuvent jouer le rôle d'intermédiaire pour ces attributs discriminants. Par exemple, le code ZIP est un intermédiaire de l'origine ethnique aux États-Unis. La dernière approche elle aussi est confrontée au problème des attributs intermédiaires, en outre la complexité du modèle global est plus élevée, car on double le nombre de modèles à entraîner et cette complexité croîtrait linéairement avec le nombre d'attributs jugés discriminants. Enfin des questions de légitimité peuvent être posées dues au traitement spécifique appliqué à chaque groupe. En effet, on pourrait trouver deux profils quasiment identiques appartenant chacun à un groupe respectif, mais pour lesquels les décisions sont différentes dues aux différents classifieurs utilisés. Ces exemples démontrent la nécessité d'analyse et de transparence de ces algorithmes de personnalisation, de mettre sur pieds des outils permettant aux usagers (concepteur, développeur et régulateur), de vérifier et de certifier la qualité du système développé.

La suite de ce document sera découpée en trois grandes sections : la première traitant de la discrimination où nous présenterons les définitions d'équité, les techniques et méthodes pour réaliser des systèmes non discriminants (chapitres 1 et 2), la seconde concernera la transparence des algorithmes, nous y aborderons en particulier les notions et les outils de transparence (chapitre 3) ; dans la dernière, nous présenterons les travaux que nous prévoyons de réaliser durant le reste de cette thèse de doctorat (chapitre **??**).

CHAPITRE I

CLASSIFIER DES DONNÉES SOUS CONTRAINTE DE L'ÉQUITÉ

Comme mentionné dans l'introduction, l'usage intensif des algorithmes de décision automatique amène avec lui des enjeux de discrimination. Dans cette section, nous commencerons par introduire la notion d'équité, en la définissant puis en décrivant des méthodes pour la garantir.

1.1 *Modèle*

Avant toute chose, nous devons d'abord définir les notations que nous utiliserons par la suite. Premièrement, nous considérerons un ensemble de données D, composé de N profils d'individus. Chaque ligne $x_i \in D$ représente l'individu i caractérisé par d attributs qui peuvent être soit catégoriels, discrets ou continus. Ces attributs peuvent être divisés en trois catégories : les attributs sensibles ou protégés S, les attributs de décision Y et le reste des attributs A. Les attributs *sensibles* ou *protégés* S correspondent aux attributs pouvant conduire à la discrimination et légalement exclus du cadre d'un processus de décision. Par exemple une décision ne peut être fondée sur les convictions religieuses, le genre ou l'origine ethnique d'un individu. Dans la suite de ce document, nous ne considérerons qu'il n'existe qu'un seul attribut jugé sensible dans le jeu de données et cet attribut est binaire (c'est-à-dire dont le domaine n'est constitué que de deux valeurs). Cependant, il est toujours possible de transformer un attribut multivalué en attribut binaire en regroupant ses valeurs en deux ensembles disjoints. Toutefois,

une attention particulière devra être portée aux phénomènes statistiques tels que le paradoxe de Simpson [1] [Malinas et Bigelow, 2004] qui peuvent apparaître et conduire à des interprétations erronées des résultats. À partir des valeurs de cet attribut, la population globale peut-être divisée en deux groupes distincts :

— Le *groupe protégé* S_0 : $S = 0$. Il s'agit du groupe qui est le plus susceptible d'être discriminé, tel que les minorités ethniques. Il est parfois appelé groupe minoritaire.

— Le *groupe par défaut* S_1. Souvent appelé groupe majoritaire dans la littérature, il s'agit du groupe pour lequel $S = 1$.

L'*attribut de décision* ou *décision* Y correspond par exemple à la prédiction d'un classifieur (accepté ou rejeté par exemple pour une candidature). Il peut ne pas être présent dans le jeu de données. De même que pour l'attribut sensible, nous supposerons pour le reste de ce document qu'un seul attribut de décision est utilisé, et cet attribut est binaire.

Les autres attributs A, qui ne sont ni sensibles, ni des décisions seront considérés comme non sensibles. En protection de la vie privée, ces attributs parfois sont appelés *quasi-identifiants*, car leur combinaison permet souvent d'identifier de manière unique les individus (par exemple, la combinaison du code postal, le genre et la date de naissance).

1.2 *Définition et quantification de l'équité*

La discrimination et l'équité sont des concepts opposés. En effet, tandis que la discrimination représente le fait de traiter injustement une personne ou une population, l'équité correspond à l'égalité des traitements. Il existe plusieurs définitions du concept d'équité dans la littérature, et certaines d'entre elles sont mutuellement exclusives. Nous allons par la suite discuter de trois principales.

---

1. Paradoxe de Simpson : paradoxe statistique dans lequel phénomène observé de plusieurs groupes semble s'inverser lorsque les groupes sont combinés.

1.2.1 Équité de groupe

L'équité de groupe [Feldman *et al.*, 2015 ; Barocas et Selbst, 2016], aussi appelée *parité statistique* consiste à ce que les statistiques de décisions (taux d'acceptation ou de refus de prêt par exemple) dans les deux groupes soient les mêmes entre le groupe minoritaire et le groupe par défaut. L'équité de groupe est liée à la notion de "*Disparate Impact*", qui apparaît lorsque le ratio entre la proportion de décisions positives dans le groupe par défaut et celle dans le groupe protégé est en deçà d'un certain seuil. L'équité de groupe, comme son nom l'indique, est mesurée en rapport avec les partitions de population définies par les valeurs de l'attribut sensible. L'équité de groupe peut ainsi être quantifiée de plusieurs manières :

— Soit en considérant le ratio des décisions positives dans les deux groupes :

$$DispImpactRatio = \frac{Pr(Y = 1|S_1)}{Pr(Y = 1|S_0)} \tag{1.1}$$

— Soit en considérant la différence de décisions positives dans les groupes protégé et par défaut :

$$DispImpact = |Pr[Y = 1|S_1] - Pr[Y = 1|S_0]| \tag{1.2}$$

Dans le premier cas, la valeur optimale est de 1 alors que dans le second la parité statistique existe si la métrique vaut 0. Il existe d'autres métriques utilisées dans la littérature (ex. *rapport de chances (odds ratio), extended lift,* etc.), mais elles sont fortement corrélées à celles présentées ci-dessus. Notons toutefois que certaines approches quantifient la discrimination par rapport aux décisions négatives ($y = 0$) en utilisant les mêmes formulations que celles sus-citées. Le retrait de la valeur absolue dans l'équation 1.2 permet de quantifier explicitement la direction dans laquelle la discrimination s'effectue.

Contrairement à l'équité individuelle (définie dans la section suivante) qui peut être liée au concept de méritocratie, l'équité de groupe soulève des questions de légitimité.

En effet, elle impose des traitements similaires à une échelle macroscopique ; ainsi, le respect de cette propriété peut nécessiter l'acceptation (respectivement le rejet) de personnes qui à priori ne devrait pas l'être, ce qui a pour conséquence de ne pas respecter l'équité individuelle. Par exemple, en considérant un jeu de données constitué de 20% de profils appartenant au groupe par défaut, si 90% de ces profils ont des décisions positives, l'équité de groupe implique que 90% des profils du groupe protégé reçoivent aussi une décision positive. Dans ce contexte, l'équité individuelle n'est respectée que si les profils aux décisions positives sont tous similaires, soit 90% des profils du jeu de données sont similaires. Puisqu'il serait difficile d'obtenir une telle proportion de similarité, l'obtention d'une égalité des taux impliquerait donc que des profils non similaires obtiennent la même décision et par extension, des profils qui devraient être à priori rejetés se verront donc attribuer des décisions positives.

La notion d'*égalité des chances (equality of odds)* introduite par [Hardt *et al.*, 2016] étend le concept d'équité de groupe en imposant que les profils aient les mêmes chances d'être acceptés véritablement ou par erreur :

$$Equality\ of\ odds : Pr(\hat{Y} = 1|S_0, Y = y) = Pr(\hat{Y} = 1|S_1, Y = y), y \in \{0, 1\} \qquad (1.3)$$

($\hat{Y}$ est la décision d'un classifieur sur un profil particulier). Autrement dit, on impose l'égalité des proportions de vrais positifs et de faux positifs entre les deux groupes. Ainsi, le classifieur ne favorisera aucune des sous-populations. L'équité de groupe pourrait être vue comme un relâchement de l'égalité des chances. En posant $y = 1$, on obtient *l'égalité d'opportunités (equality of opportunity)*.

1.2.2 Équité individuelle

L'équité individuelle (*individual fairness*) [Dwork *et al.*, 2012] reflète une notion naturelle selon laquelle des individus aux profils similaires devraient recevoir la même décision. L'application de cette propriété dans notre contexte conduit à ce que des individus aux profils proches, à l'exception de l'attribut sensible, reçoivent la même

décision. Cette notion fait référence au concept de *«disparate treatment»* [Barocas et Selbst, 2016], qui existe lorsque l'attribut sensible est utilisé (directement ou indirectement) dans le processus de décision.

[Zemel *et al.*, 2013] ont introduit la métrique de *cohérence des décisions* pour quantifier cette équité. Étant donné le voisinage $kNN(x_n)$ d'un profil x_n et la décision y_n associée au profil x_n :

$$yNN = 1 - \frac{1}{Nk} \sum_{n=1} |y_n - \sum_{j \in kNN(x_n)} y_j| \tag{1.4}$$

Ainsi, cette métrique permet de capturer la dissimilarité entre les décisions attribuées à des profils proches selon la définition de la métrique de similarité que l'on choisit. Il est important de noter qu'obtenir une cohérence de $yNN = 1$ est impossible même pour un système parfaitement consistent, à moins d'avoir la même décision pour tous les individus. L'objectif est donc d'obtenir $0.5 < yNN \leq 1$. Autour de 0.5, pour chaque individu, on peut trouver dans son voisinage autant de décisions de chaque classe - la surface de décision est complexe et hétérogène.

1.2.3 Équité par l'impossibilité d'inférer l'attribut sensible

Nous pouvons aussi définir l'équité comme étant l'inutilisation (directe ou indirecte) de l'attribut sensible dans le processus de décision (*fairness through unawareness*)[Verma et Rubin, 2018]. L'idée principale qui régit cette notion d'équité est le fait qu'il est impossible de discriminer si l'on ne connaît à quel groupe un profil appartient. Ainsi, il faut masquer le mieux possible l'attribut sensible, en rendant l'inférence de l'attribut sensible difficile pour tout attaquant qui essaie de reconstruire cet attribut à partir d'autres informations disponibles dans le jeu de données ou de connaissances externes.

Considérons un classifieur $f : \{A, Y\} \rightarrow S$ dont la tâche serait de reconstruire S à partir de $\{A, Y\}$. Pour quantifier la protection de l'attribut discriminant, une métrique

standard est l'exactitude de la prédiction de S (*sensitive accuracy*) $sAcc$ [Zemel *et al.*, 2013] :

$$sAcc(f(\{A, Y\}), S) = 1 - \frac{1}{N}\sum_{i=1}^{N} |f(\{a_i, y_i\}) - s| \in [0 - 1] \tag{1.5}$$

$sAcc$ varie entre 0 et 1, et la valeur optimale est de 0.5 lorsque S est binaire et qu'il y a autant de profil dans le groupe protégé que dans le groupe par défaut (groupes équilibrés). En effet, un taux de 0.5 signifie que le classifieur n'est pas meilleur qu'un choix aléatoire. Lorsque les groupes sont déséquilibrés, un classifieur aléatoire aurait un taux de prédiction correcte proche de la proportion de la classe majoritaire P_{maj}. Un classifieur ayant un taux de prédiction correcte entre 0.5 et P_{maj} (ou P_{min} et 0.5) pourrait donc être considéré comme inefficace pour reconstruire l'attribut sensible.

L'inaptitude de f à prédire S peut aussi être aussi quantifié par le *BER (Balanced Error Rate)* [Feldman *et al.*, 2015].

$$BER(f(A, Y), S) = \frac{1}{2}[\frac{\sum_{n:s_n=1} f(a_n, y_n) = s_0}{\sum_{n:s_n=1} 1} + \frac{\sum_{n:s_n=0} f(a_n, y_n) = s_1}{\sum_{n:s_n=0} 1}] \tag{1.6}$$

Le BER capture la qualité de prédiction de f dans chacun des groupes, en mesurant respectivement les taux de faux positifs dans chaque groupe. La valeur optimale est de 0.5 qui veut dire que f n'a pas une meilleure performance qu'un classifieur aléatoire, ou qu'un classifieur qui prédirait toujours la classe majoritaire.

1.3 *Atteindre l'équité par conception*

Dans la section précédente, nous avons vu comment définir et quantifier l'équité. Plusieurs techniques ont été développées afin de limiter les risques de discrimination, elles peuvent être regroupées en plusieurs familles [Friedler *et al.*, 2018]. On distingue ainsi :

— Les techniques de *prétraitement* des données [Ruggieri, 2014 ; Kamiran et Calders, 2012 ; Luong *et al.*, 2011 ; Zhang *et al.*, 2016 ; Feldman *et al.*, 2015 ; Zemel *et al.*, 2013 ; Louizos *et al.*, 2015 ; Edwards et Storkey, 2015 ; Feldman *et al.*,

2015 ; Zemel *et al.*, 2013 ; Louizos *et al.*, 2015 ; Edwards et Storkey, 2015]. Pour ces techniques, l'idée sous-jacente est de retirer les biais existant envers certains groupes de la population afin de pouvoir réaliser une classification non discriminatoire. On distingue deux types de prétraitement : le prétraitement optimisé pour une tâche bien spécifique, qui a pour but de mitiger la discrimination en changeant l'espace de représentation tout en préservant l'utilité par rapport à cette tâche précise, et celui optimisé pour introduire le moins possible de perturbations (même espace que les données originales). Pour cette dernière les données peuvent ensuite être utilisées pour n'importe quelle tâche subséquente. Une particularité du prétraitement est qu'il peut-être fait en utilisant des algorithmes d'apprentissage automatique.

— Les techniques de *modification des algorithmes d'apprentissage* [Calders *et al.*, 2013 ; Bilal Zafar *et al.*, 2017 ; Kamishima *et al.*, 2012 ; Dwork *et al.*, 2012]. Ici, on introduit des modifications dans l'algorithme d'apprentissage de sorte que les critères d'équité soient respectés par construction. Les modifications apportées sont souvent introduites par une méthode de régularisation. Ces méthodes présentent pour la plupart l'inconvénient d'être spécifique à un certain type de classifieur.

— Les techniques de *post-traitement des données* [Kamiran et Calders, 2009 ; Kamiran *et al.*, 2010 ; Hardt *et al.*, 2016]. Ces approches consistent à modifier les résultats du classifieur déjà entraîné, de sorte que l'on obtienne les résultats souhaités. Ces méthodes ont pour avantage d'être directement applicable sur des classifieurs déjà entraînés. En effet, pour certaines applications, l'entraînement du classifieur requiert des ressources conséquentes et nécessite très souvent un temps et un volume de données considérable. De ce fait, assurer l'équité sans avoir à réentraîner l'ensemble du système se révèle être un avantage important. Tout comme les approches par modification d'algorithmes, certaines de ces approches sont spécifiques au type de classifieur utilisé.

Dans le reste de cette section, nous aborderons dans l'ordre les méthodes associées à l'équité de groupe, celles associées à l'équité individuelle et celles associées à l'équité par prévention d'utilisation d'attribut discriminant.

1.3.1 L'équité de groupe

Prétraitement. [Ruggieri, 2014] propose une adaptation de la *t-proximité* [Ninghui *et al.*, 2007] pour contrôler le niveau de discrimination. Étant donné S et A, on définit par classe d'équivalence les partitions non vides de la population formée par la combinaison des différentes valeurs de A. Par exemple, si l'on considère trois attributs non sensibles $A_i, i \in \{1,2,3\}$ ($a_i \in A$) ayant pour domaines respectifs $dom(A_i), i \in \{1,2,3\}$, les classes d'équivalence seraient formées par les profils qui *supportent* chacun des ensembles formés par les combinaisons attributs-valeurs $< A_1 = a_1, A_2 = a_2, A_3 = a_3 >$, $a_i \in dom(A_i)$. Ces ensembles sont appelés ensembles d'objets et l'ensemble des profils qui les *supportent* constituent leur couverture. Dans la littérature en protection de la vie privée, les attributs A_i sont appelés *quasi-identifiants*, et une classe d'équivalence est formée par tous les profils dont les quasi-identifiants ont les mêmes valeurs. La *t-proximité* est satisfaite lorsque la distribution d'un attribut sensible dans n'importe quelle classe d'équivalence est très proche de la distribution de cet attribut dans la base de données de manière générale : $|Pr[S_0]_{eq.class} - Pr[S_0]_{table}| \leq t$. Dans ce cas, on dit que la classe d'équivalence est *t-fermée*, où la valeur de t représente la distance maximale que l'on est prêt à tolérer. La table est *t-fermée* si toutes les classes d'équivalence sont *t-fermées*. Salvatore Ruggieri a démontré que si un jeu de données est *t-fermé*, alors la discrimination est bornée par une valeur qui elle aussi dépend de t :

$$DispImpact \in [0, min\{2t, t + \hat{p}, 1\}], \hat{p} = min\{Pr[Y = 0], Pr[Y = 1]\} \tag{1.7}$$

$$DispImpactRatio \in [\frac{max\{Pr[Y = 1] - t, 0\}}{min\{Pr[Y = 1] + t, 1\}}, 1] \tag{1.8}$$

Une adaptation de l'algorithme de t-proximité a été proposée dans le cadre de cette approche. Puisque cette approche est issue de la t-proximité, elle présente l'avantage

de généraliser les données au lieu de les perturber. Cependant, l'approche souffre des mêmes limites que la t-proximité, notamment le fait que l'on ne peut pas généraliser les données plus que la table elle-même. En effet, la généralisation consiste à regrouper des valeurs d'attributs, jumelant ainsi plusieurs partitions en une seule. Puisque la table n'est pas une partition, il n'est pas possible de la généraliser. Ainsi, changer la valeur de la mesure de discrimination revient à perturber la distribution de l'attribut de sensible (ou celui de décision), ce qui est contraire au principe des méthodes de généralisation.

[Kamiran et Calders, 2012] ont proposé plusieurs procédures de prétraitement des données pour prévenir la discrimination, notamment le massage, la pondération et l'échantillonnage des données. Le massage des données consiste tout simplement à promouvoir et rétrograder un nombre M de candidats dans le jeu de données d'entraînement. Ce nombre est proportionnel au degré de discrimination existant : $M = N * DispImpact * Pr[Y = 0] * Pr[Y = 1]$. Les candidats à la promotion sont sélectionnés parmi les profils du groupe protégé ayant des décisions négatives ($pro_{group} = \{S_0 \cap \{Y = 1\}\}$), tandis que les candidats à la rétrogradation sont pris dans l'ensemble des profils ayant des décisions positives dans le groupe par défaut ($ret_{group} = \{S_1 \cap \{Y = 1\}\}$). Les profils sont ordonnés suivant leur probabilité d'acceptation (probabilité d'obtenir une décision positive) obtenue par un classifieur tel quun classifieur bayésien. L'ordonnancement se fait dans l'ordre décroissant dans pro_{group} et dans l'ordre décroissant dans ret_{group}. Enfin, les décisions du top-M de ces deux groupes sont inversées, l'idée étant de modifier uniquement les profils les plus proches de la frontière de décision, car ils sont les plus susceptibles de changer de classes.

La pondération consiste à rajouter un attribut de *poids* dans le jeu d'entraînement, cet attribut sera aussi pris en compte en entrée du classifieur pour la tâche. Le but de la pondération est de compenser la discrimination en assignant des poids plus importants aux profils protégés ayant des décisions positives qu'à ceux ayant des décisions négatives et en réduisant l'importance des profils positifs du groupe par défaut en leur assignant

une pondération inverse de celle du groupe protégé. Le poids d'un profil x_i est calculé par : $w_i = \frac{Pr[S = s_i] * Pr[Y = y_i]}{Pr[Y = y_i | S = s_i] * Pr[S = s_i]}$. Le numérateur représente le cas idéal de non-discrimination où S et Y sont indépendants (auquel cas $w_i = 1$ pour tous les profils x_i), le dénominateur représente à la distribution observée.

L'échantillonnage choisit aléatoirement un certain nombre d'individus dans les groupes protégés et majoritaires, partitionnés de plus en fonction des valeurs de l'attribut de décision (4 groupes). On choisit respectivement $M_{s_0}^{y_0}$, $M_{s_0}^{y_1}$, $M_{s_1}^{y_0}$, $M_{s_1}^{y_1}$ dans les groupes protégé aux décisions négatives, protégé aux décisions positives, par défaut aux décisions négatives et par défaut aux décisions positives. Le nombre $M_{s_j}^{y_i}$ est déterminé de la même manière que les poids w_i : $M_{s_j}^{y_i} = \frac{Pr[Y = y_i]}{Pr[Y = y_i | S = s_j]}$. En tenant compte du fait que les profils les plus proches de la frontière de décision sont les plus susceptibles d'être discriminés, [Kamiran et Calders, 2012] proposent aussi un échantillonnage qui combine la pondération des profils.

Modification d'algorithme. Pour contrôler la discrimination en régression linéaire, [Calders *et al.*, 2013] modélisent la métrique d'équité de groupe comme une contrainte de la fonction objectif. Soit x_i/y le profil x_i auquel on a retiré la décision. L'approche consiste à :

$$\begin{aligned} &Minimiser && L(\theta) = (\theta * x_i/y - y_i)^2 \\ &sous\ contrainte\ que && \frac{\sum_{n:s_n=1} \theta * x_n/y}{\sum_{n:s_n=1} 1} = \frac{\sum_{n:s_n=0} \theta * x_n/y}{\sum_{n:s_n=0} 1} \end{aligned}$$

Avec thêta les paramètres de la régression linéaire. En utilisant la formulation duale, on peut introduire la contrainte en tant que régularisation et résoudre l'équation. Ainsi, l'objectif est d'apprendre une modélisation dont les paramètres sont indépendants du groupe d'appartenance, et qui permet de prédire correctement l'attribut de décision.

De manière plus générale que [Calders *et al.*, 2013], [Bilal Zafar *et al.*, 2017] proposent de contrôler la discrimination pour tout classifieur convexe utilisant une frontière de

décision. L'approche consiste à minimiser l'erreur d'entraînement $L(\theta)$ sous contrainte que la covariance entre l'attribut sensible S et la distance d'un profil par rapport à la frontière de décision $d_\theta(x_i)$ ($Cov(S, d_\theta(x))$) soit comprise dans l'intervalle $[-c, c]$. En posant $\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i$ et en observant que $E[(s_i - \bar{s})] = 0$, cela correspond à :

$$\begin{aligned} &\textit{Minimiser} && L(\theta) \\ &\textit{sous contrainte que} && |\frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s}) * d_\theta(x_i)| \leq c \end{aligned}$$

La covariance mesure la probabilité conjointe entre S et d_θ et donc entre S et Y. Le fait de borner cette valeur permet ainsi de limiter la dépendance qui existe entre ces deux variables. Le paramètre c permet de contrôler le compromis entre le taux de prédiction correcte et l'équité de sorte que pour une valeur de c très proche de zéro, on obtient $Pr(d_\theta \geq 0|S_0) = Pr(d_\theta \geq 0|S_1)$ (équité de groupe), mais au prix d'un taux de prédiction correcte amoindri. La fonction objectif ci-dessus correspond à la maximisation du taux de bonnes prédictions sous contrainte de l'équité. On peut aussi exprimer le problème connexe qui est la maximisation de l'équité sous contrainte du taux de bonnes prédictions :

$$\begin{aligned} &\textit{minimiser} && |\frac{1}{N}\sum_{i=1}^{N}(s_i - \bar{s}) * d_\theta(x_i)| \\ &\textit{sous contrainte que} && L(\theta) \leq (1+\gamma)L(\theta^*) \end{aligned}$$

$L(\theta^*)$ représente l'erreur d'entraînement optimal sur l'ensemble d'entraînement, γ permet de contrôler la tolérance que l'on accepte au niveau du taux de prédictions justes. Tout comme [Calders *et al.*, 2013], pour ces deux objectifs, les contraintes peuvent être exprimées sous forme de coefficient de régularisation en utilisant la forme duale du problème.

Post-traitement. [Kamiran et Calders, 2009] résolvent la problématique de discrimination en inversant les décisions prises par le classifieur bayésien naïf. Cette méthode

correspond notamment au massage des données présentées dans [Kamiran et Calders, 2012] (cf. section 1.3.1), cependant, le massage est appliqué en fin de classification, après que les probabilités d'acceptation ont été produites par l'algorithme. Les mêmes auteurs proposent un post-traitement des arbres de décisions pour assurer l'équité [Kamiran *et al.*, 2010]. Leur méthode consiste à inverser les décisions dans un certain nombre de feuilles pour réduire la discrimination plutôt que de considérer la classe majoritaire pour assigner la décision. Chaque inversion de décision réduit le taux de prédiction correcte de Δ_{acc}, et diminue la discrimination de Δ_{disc}. Il faut donc trouver la combinaison optimale du nombre de feuilles à inverser, chacune avec son poids $-\Delta_{disc}$ et son profit $-\Delta_{acc}$. Résoudre ce problème revient ainsi à résoudre le problème du *sac à dos* [Martello, 1990], qui est un problème très connu dans le domaine de l'optimisation.

Les auteurs ont aussi proposé une modification de l'arbre de décision dans laquelle la sélection des noeuds pour effectuer la division de l'arbre se fait selon un critère qui minimise la discrimination, au lieu de maximiser le taux de bonnes prédictions. Par exemple, un critère d'optimisation utilisé est le gain d'information par rapport à la décision (IGC). Les auteurs appliquent le gain d'information par rapport à l'attribut sensible (IGS), et combinent ces gains pour établir un compromis taux de bonnes prédictions-discrimination. La combinaison peut prendre les formes $IGS + IGC$, $IGS - IGC$ ou $\frac{IGS}{IGC}$.

1.3.2 L'équité individuelle

Une difficulté inhérente à l'équité individuelle est la notion de similarité entre les profils, qui est étroitement liée à la détection de la discrimination. Nous n'avons pas été en mesure de trouver des méthodes de post-traitement des données pour cette métrique d'équité. Il est important de noter que la non-utilisation de l'attribut sensible en entrée du classifieur n'est pas suffisante pour garantir l'équité individuelle. Ci-après nous allons passer en revue quelques approches pour atteindre l'équité individuelle.

Prétraitement. [Luong *et al.*, 2011] ont développé une approche inspirée du test de situation, utilisé dans le cadre légal pour découvrir la discrimination. Le test de situation consiste, lorsqu'une plainte est déposée pour discrimination envers un individu sur un processus de décision (un recrutement par exemple), à faire participer au même processus des candidats *tests* qui ont les mêmes caractéristiques que l'individu concerné à l'exception des caractéristiques discriminantes. Un profil du groupe protégé x a donc un ensemble Ω de profils protégés similaires ($S = s_0$) et un ensemble Ψ de profils similaires appartenant au groupe par défaut ($S = s_1$). L'objectif est de mesurer la dissimilarité entre les décisions obtenues dans les deux groupes. [Luong *et al.*, 2011] automatise la sélection d'individus similaires en utilisant une moyenne de 3 distances entre attributs : la distance de *Manhattan* pour les attributs numériques (préalablement standardisés), le *pourcentage de discordance* pour les attributs catégoriques, et la *différence de rang normalisée* pour les attributs ordinaux. La discrimination peut ainsi être mesurée par la différence de proportions de décisions négatives entre les groupes Ω et Ψ. Une valeur négative signifie la présence de discrimination, tandis qu'une valeur positive reflète du favoritisme envers la population protégée. Remédier à la discrimination se fait en inversant la décision du profil x uniquement si x appartient au groupe protégé, sa décision est négative et la discrimination mesurée pour ce profil x est supérieure à un seuil fixé. Les données sont ainsi prétraitées avant d'être utilisées pour la classification. [Zhang *et al.*, 2016] étendent ce cadre en n'utilisant que les attributs qui ont une influence directe sur la décision pour sélectionner les profils similaires. Leur approche utilise un *réseau bayésien causal* pour raisonner sur les causalités et les inférences afin de sélectionner les attributs pertinents.

Modification d'algorithme. [Dwork *et al.*, 2012] formalisent l'équité individuelle comme étant une condition de Lipschitz sur le classifieur. Étant donné deux profils x_i, x_j, les probabilités associées $M(x_i, y = b)$, $M(x_j, y = b)$ ($b \in \{0, 1\}$) obtenues à la

sortie d'un classifieur pour chacun des profils respectifs, le classifieur satisfait l'équité individuelle (propriété $(D, d) - Lipschitz$) si

$$D(M(x_i, y = k), M(x_j, y = k)) \leq d(x_i, x_j) \tag{1.9}$$

D est une mesure de distance entre distributions, tandis que d est une mesure de similarité qui est définie selon l'application et est souvent utilisée de manière implicite ou explicite (par exemple : "Les personnes qui achètent X et dont le code postal est Z sont similaires aux personnes qui habitent T et achètent K", etc.).

La classification revient ainsi à trouver une modélisation de la distribution des profils vers la distribution des décisions (Y) qui minimise la perte en utilité, c'est-à-dire l'erreur de classification sous contrainte de la condition de Lipschitz.

Une approche située entre l'équité individuelle et l'équité par prévention d'utilisation de l'attribut sensible est celle développée par [Kamishima *et al.*, 2012], qui consiste à introduire la contrainte d'équité sous forme de terme de régularisation dans le processus d'entraînement du classifieur. La forme finale de la fonction objectif est donc obtenue en combinant l'erreur d'entraînement, une première régularisation pour éviter le sur-apprentissage et une seconde pour assurer l'équité. Ce second terme de régularisation, appelé *indice de préjudice* dans leur approche correspond à l'information mutuelle entre Yet S, caractérisant de ce fait la dépendance des décisions à l'attribut sensible. On obtient en fin d'entraînement, un classifieur capable de prendre des décisions indépendamment de l'attribut sensible (*disparate treatment* - équité individuelle), car l'attribut sensible n'est pas connu ou difficile à inférer.

1.3.3 L'équité par prévention d'utilisation de l'attribut sensible

Afin d'assurer cette propriété, les techniques développées appartiennent en général à la famille de méthodes de prétraitement des données, car le but est de pouvoir masquer toute information relative à l'attribut sensible avant d'effectuer des calculs sur l'en-

semble de données. Les techniques que l'on présentera par la suite essayent toutes de réduire les corrélations qui existent entre S et les autres attributs afin de prévenir sa reconstruction et son utilisation dans le processus de décision.

Pour masquer l'attribut sensible dans le jeu de données, [Feldman *et al.*, 2015] proposent de transformer les distributions de valeurs des attributs pour les groupes protégés et par défaut en la distribution médiane. Plus précisément, étant donné l'ensemble des attributs A, la distribution médiane est obtenue en construisant dans un premier temps la fonction de répartition (CDF) de ces attributs, conditionnée par leur groupe d'appartenance - $F(A|S=0)$ et $F(A|S=1)$- à partir des probabilités $P(A|S=0)$ et $P(A|S=1)$. La distribution médiane est ainsi obtenue en calculant la médiane des fonctions de répartition : $F_{median}(F(A|S=0), F(A|S=1))$. Le choix de la médiane est justifié par le fait que celle-ci minimise la distance entre les deux distributions, introduisant ainsi le minimum de perturbations tout en assurant un bon niveau de protection de S. Pour assurer un compromis entre la perturbation et la protection, un paramètre λ est introduit, permettant de décider à quel point l'on veut rapprocher les distributions de la médiane. Pour tenir compte des corrélations plus complexes (impliquant plus d'un attribut par exemple), les auteurs utilisent un classifieur dont la tâche est de prédire l'attribut sensible. L'incapacité du classifieur à réussir la tâche constitue une garantie de protection. Cependant, nous pensons que le dommage introduit par cette approche est important, car chaque attribut est modifié indépendamment des autres.

[Zemel *et al.*, 2013] ont développé une méthode pour apprendre une nouvelle représentation des données dans laquelle la contrainte d'équité est respectée. Cette approche repose sur un ensemble de prototypes Z - qui peuvent être vus comme des profils représentatifs et diversifiés du jeu de données, à partir desquels l'ensemble des profils seront construits. Ainsi, un profil x est représenté par une combinaison linéaire des différents prototypes $z_k \in Z : x_i = \sum_k p_k^i * z_k$. p_k^i correspond à la contribution du prototype k dans le profil i. L'attribut sensible est masqué en s'assurant que l'espérance de la contribution

de chaque prototype dans le groupe protégé et dans le groupe par défaut soit la même : $|\frac{\sum_{i \in S_0} p_k^i}{|S_0|} - \frac{\sum_{i \in S_1} p_k^i}{|S_1|}| = 0$. Les décisions Y sont prises à partir d'un classifieur qui utilise les probabilités p_k^i : $\hat{y}_i = \sum_k p_k^i \omega_k$. L'objectif de cette approche est d'apprendre la distribution de probabilités $P(Z|x)$ de sorte que $P(Z|x, S = 0) = P(Z|x, S = 1)$, sous contrainte que la reconstruction du profil x soit la meilleure possible ($\sum_k p_k^i * z_k - x = 0$) et qu'un classifieur utilisé pour une tâche ait le meilleur taux de prédiction correcte, $|\hat{Y} - Y| = 0$. Pour cette méthode, le choix et le nombre de prototypes sont cruciaux pour son bon fonctionnement.

[Louizos *et al.*, 2015] ont montré que les auto-encodeurs variationnels [Kingma et Welling, 2013] (VAE) incorporent déjà la notion d'équité de groupe dans leur définition. L'encodeur dans ce modèle est contraint à produire des données de faible dimension qui doivent suivre une distribution préalablement choisie indépendamment de l'attribut protégé. Pour donner plus de détails, un auto-encodeur [Baldi et Hornik, 1989] est un modèle qui consiste en deux réseaux de neurones, l'encodeur et le décodeur. L'encodeur produit une représentation des données Z à partir des entrées, tandis que le décodeur reconstruit ces entrées à partir de Z. Les deux modèles sont contraints par une erreur de reconstruction. Les auto-encodeurs variationnels fonctionnent sur le même principe que les auto-encodeurs à l'unique différence que Z a été préalablement choisi (une gaussienne par exemple), l'encodeur devant ainsi produire une représentation qui suit cette distribution. Ainsi puisque l'on choisit la distribution de Z indépendamment de S, l'encodeur produit une représentation qui transforme les distributions des groupes protégé et par défaut vers la même distribution finale. Toutefois, si les profils du groupe protégés sont très différents des profils du groupe par défaut, on pourra observer des groupements de points dans la distribution Z qui permettront de ré-identifier le groupe d'appartenance des profils. Les auteurs introduisent la contrainte d'équité selon laquelle chaque point de Z doit avoir autant de chance d'identifier chacun des deux groupes, le modèle final est appelé *Auto-Encodeurs variationels équitable (VFAE)*. Cette approche

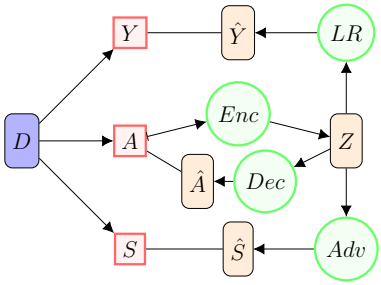


Figure 1.1 : Modèle ALFR. LR : Régression Logistique, Enc : Encodeur, Dec : Decodeur, Adv : Adversaire.

est similaire à celle de [Zemel *et al.*, 2013] dans la mesure où l'on cherche à cacher le groupe d'appartenance d'un profil en apprenant une représentation dans laquelle chaque point a autant de chance de provenir de chacun des deux groupes. Afin de préserver les décisions du jeu de données, l'approche a été modifiée de sorte que l'objectif final soit d'apprendre une représentation Z_2 qui lorsqu'elle est combinée à la distribution des décisions Y, produit la représentation Z. Le dommage dans cette approche ne peut être mesuré que par l'utilisation de l'attribut de décision.

Une des approches les plus récentes pour assurer l'équité est l'approche *ALFR* conçue par [Edwards et Storkey, 2015]. Ici tout comme dans les méthodes précédentes, l'objectif est aussi d'apprendre une représentation qui est équitable pour une tâche précise. L'approche se rapproche d'une combinaison des trois approches précédentes, dans la mesure où elle utilise un classifieur pour une tâche afin de préserver l'utilité des données [Zemel *et al.*, 2013], la garantie d'équité étant donnée par l'impossibilité d'un adversaire à prédire S [Feldman *et al.*, 2015], tout en garantissant que l'on puisse reconstruire les données originales à partir de la représentation apprise [Louizos *et al.*, 2015]. La figure 1.1 illustre l'approche, le classifieur utilisé pour cette tâche ici étant ici la régression logistique.

De même que pour les VFAE, le dommage introduit dans les données est difficile à

quantifier, car la représentation obtenue n'est pas interprétable. La plupart de ces approches de prétraitement pour prévenir l'utilisation de l'attribut protégé en classification reposent sur le changement d'espace des données, de l'espace original vers un autre dans lequel la propriété d'équité est respectée. Ce changement d'espace soulève des questions d'interprétation des décisions qui seraient prises à partir de ces nouvelles représentations (en particulier on peut se demander par rapport à quels critères la décision a été prise). Ces questions sont notamment très importantes dans le contexte où la transparence et l'interprétabilité sont de plus en plus demandées. De même, la représentation trouvée est celle qui étant la mieux adaptée pour une tâche spécifique, les données ainsi assainies ne peuvent pas être facilement utilisées pour une autre tâche. La préservation de la distribution de l'attribut de décision soulève aussi des questions d'équité dans la mesure où ces décisions peuvent être issues d'un processus discriminant. Ainsi préserver ces décisions pourrait revenir à préserver la discrimination. Pour pallier à ces situations, nous avons développé une approche inspirée des *réseaux adversariaux génératifs (RAGS)*. Notre approche présente en plus l'avantage de préserver les données assainies dans le même espace que celui des données originales, ce qui garantit donc une interprétabilité de la transformation effectuée et la compréhensibilité du jeu de données est préservée.

CHAPITRE II

ASSAINISSEMENT DES DONNÉES EN VUE D'AMÉLIORER L'ÉQUITÉ PAR DES RÉSEAUX ADVERSARIAUX GÉNÉRATIFS

Avant de rentrer dans les spécificités de notre approche, nous allons d'abord présenter ce que sont les réseaux adversariaux génératifs.

2.1 *Réseaux adversariaux génératifs (RAGs)*

Il s'agit d'une nouvelle approche développée en apprentissage automatique dans laquelle deux réseaux de neurones entrent en compétition dans un jeu à somme nulle [Goodfellow *et al.*, 2014]. Le premier réseau est appelé *générateur*, il a pour objectif d'apprendre la distribution à partir de laquelle un ensemble de données a été généré. L'objectif des RAGs est d'entraîner le générateur afin que ce dernier puisse produire des données qui suivent une distribution choisie. Plus précisément, le générateur apprend à générer des exemples de données qui suivent la distribution choisie, sans avoir à tout simplement reproduire à l'identique celles qui existent dans le jeu d'entraînement. Le second réseau est appelé *discriminateur* et son utilisation est justifiée par la difficulté de mesurer objectivement la qualité du générateur. La percée majeure des RAGs provient du fait qu'ils résolvent le problème de la génération de données à partir d'une distribution *à priori* inconnue. En effet, puisque la distribution reste inconnue (bien que nous disposions de quelques points appartenant à cette distribution, ces points constituant le jeu d'entraînement), il est impossible d'évaluer la qualité de la distribution apprise par

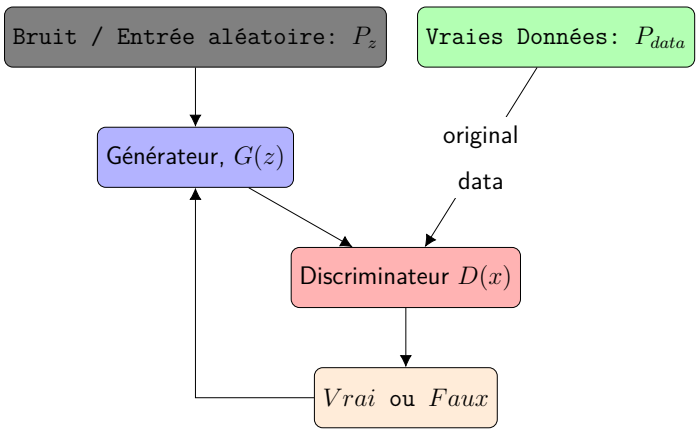


Figure 2.1 : Vue d'ensemble de la procédure des RAGs. Guidé par le discriminateur $D(x)$, le générateur $G(z)$ prend en entrée du bruit tiré de la distribution Z et produit de nouvelle données qui suivent la distribution P_{data}. Le discriminateur est entraîné à différentier entre les vraies données tirées de P_{data} des fausses produites par le générateur en prenant en entrée des échantillons des deux distributions. P_{data} n'est pas connu d'avance, seule est connue son évaluation en quelques points (les données d'entraînement).

le générateur, et ainsi le guider dans la bonne direction. Le discriminateur entre en jeu pour pallier cette situation. En entraînant le discriminateur de sorte qu'il soit en mesure de distinguer les vraies données (du jeu d'entrainement) des fausses produites par le générateur, il devient possible d'évaluer et d'approximer la différence entre la vraie distribution et celle que le générateur apprend. Le discriminateur fournit un retour au générateur sur la qualité des données générées. La figure 2.1 illustre les RAGs. À la fin de l'entraînement, le générateur est capable de produire à partir du bruit des données qui suivent la même distribution que les données originales. Depuis leurs premières expérimentations, le nombre de projets et d'approches basés sur les RAGs pour résoudre différents types de problèmes a augmenté considérablement [Hindupur, 2018]. On peut notamment citer l'approche cycle-consistent réseaux adversariaux génératifs (CycleGan) [Zhu *et al.*, 2017], qui, lorsque le modèle arrive à convergence, a appris le

style présent dans un ensemble de peintures et est capable par la suite de générer par la suite des tableaux imitant ce style à partir d'images.

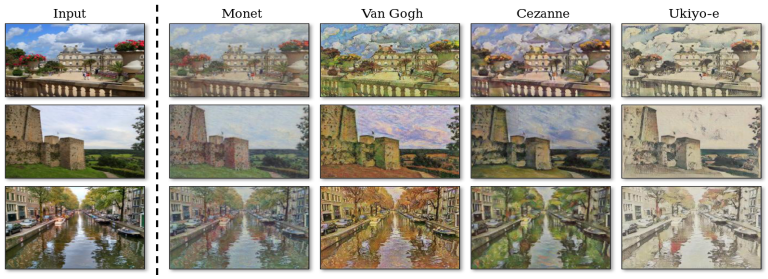


Figure 2.2 : CycleGAN. À partir des styles appris (Monet, Van Gogh, etc.), le générateur *transfert* ces styles sur les données en entrée du générateur.

Malgré leur aspect intuitif et le fait qu'ils représentent des outils performants pour modéliser des distributions, l'entraînement des RAGs peut se révéler complexe et demander très souvent beaucoup d'ingénierie pour assurer leur réussite [Zhu *et al.*, 2017]. Par exemple, durant l'entraînement, si le discriminateur ou le générateur devient beaucoup plus performant comparativement à l'autre, ce dernier ne sera pas en mesure de s'améliorer.

Notre approche GANSan est inspirée des RAGs dans la mesure où notre objectif est d'apprendre la distribution de l'attribut sensible dans le groupe protégé et le groupe par défaut, pour ensuite introduire le minimum de perturbations dans les données permettant de rendre l'inférence de l'attribut sensible difficile.

2.2 *Description de notre approche*

L'objectif de notre approche est d'entraîner un générateur que l'on appelle assainisseur, qui sera en mesure de détecter automatiquement et de mitiger toutes corrélations avec S dans le jeu de données. Ainsi il sera difficile pour un algorithme utilisant les données assainies de discriminer. De manière plus formelle, soit D l'ensemble de données original

constitué de profils $x_i = (a_i, y_i, s_i)$. On cherche à apprendre une fonction F appelée assainisseur, telle que la distance L_2 entre le jeu original et celui assaini ($L_2(D, F(D))$) soit la plus petite possible, et l'attribut protégé ne peut être reconstruit à partir de $F(D)$. F produira l'ensemble assaini $\hat{D} = F(D)$ constitué de profils $\hat{x}_i = (\hat{a}_i, \hat{y}_i, \hat{s}_i)$ et dans lequel un classifieur C ne sera pas en mesure de prédire S à partir de $\hat{A}$ et $\hat{Y}$.

$$
\begin{aligned}
L_2 &: ((A,\ Y), (\hat{A},\ \hat{Y})) \ \rightarrow \ \Re \\
&\quad ((a_i,\ y_i), (\hat{a}_i,\ \hat{y}_i)) \ \longmapsto \ L_2((A,\ Y), (\hat{A},\ \hat{Y})) \\
C &: (\hat{A},\ \hat{Y}) \ \rightarrow \ S \\
&\quad (\hat{a}_i,\ \hat{y}_i) \ \longmapsto \ s_i \\
F &: D \ \rightarrow \ \hat{D} \\
&\quad x_i \ \longmapsto \ \hat{x}_i, \quad C(\hat{a}_i,\ \hat{y}_i) \ \neq \ s_i \quad \wedge \quad min(L_2((A,\ Y), (\hat{A},\ \hat{Y})))
\end{aligned}
$$

Notons que $\hat{D}$ et D étant dans le même espace, le calcul de distance est donc possible directement. Une autre métrique de distance peut être utilisée au lieu de L_2. La procédure d'entraînement des RAGs consiste dans un premier temps à entraîner le discriminateur, puis le générateur. Le discriminateur est entraîné en premier, car c'est lui qui permet d'approximer la distribution inconnue et de guider ainsi le générateur en lui faisant savoir à quel point il produit des données éloignées de la distribution réelle qu'on cherche à apprendre. Dans notre approche, l'assainisseur doit d'abord être entraîné afin que celui-ci puisse apprendre la structure des données originales et en produire de nouvelles à partir desquelles le discriminateur sera entraîné à reconstruire l'attribut sensible. Entraîner le discriminateur en premier lieu ne serait pas utile, car l'assainisseur ne connaît pas la structure des données, et produit ainsi des données aléatoires. Le discriminateur n'est donc pas en mesure de reconstruire l'attribut sensible et la distance entre les données originales et celles produites est très grande. La procédure d'entraînement de l'assainisseur est la suivante :

— Premièrement, l'assainisseur S_{an} est entraîné. Ce dernier joue le même rôle que le générateur dans les RAGs, puisque S_{an} doit générer des données proches des données originales, mais assez différentes pour prévenir la reconstruction de S. En d'autres termes, l'assainisseur doit introduire juste assez de perturbations dans les données originales pour empêcher D_{isc} de prédire S. L'assainisseur prend en entrée les données originales D en plus d'un bruit (c'est-à-dire une valeur tirée aléatoirement) pour rendre la procédure stochastique. Cette procédure s'apparente notamment aux RAGs conditionnels [Mirza et Osindero, 2014]. Le bruit remplit deux objectifs : le premier est d'éviter le sur-apprentissage de l'ensemble dentraînement et le second est rendre la reconstruction du profil original difficile pour un attaquant.

— Deuxièmement, le discriminateur D_{isc} est entraîné à prédire l'attribut sensible à partir des données générées par l'assainisseur S_{an}. À la différence des RAGs, dans notre approche le discriminateur n'a pas connaissance des données originales. Plus le taux de prédiction correcte du discriminateur sera élevé, moins bonne sera la protection offerte par l'assainisseur, et donc le risque de discrimination sera élevé.

Ces deux étapes sont répétées durant plusieurs itérations jusqu'à convergence, auquel cas l'assainisseur produit des données semblables aux données originales, mais à partir desquelles le taux de prédictions justes d'un classifieur pour prédire S se rapproche d'un choix aléatoire. La figure 2.3 résume notre approche.

2.3 *Entraînement de GANSan*

Pour contrôler le compromis entre la quantité de dommage introduite et le niveau de protection souhaité, nous utilisons un hyper-paramètre α, qui varie entre 0 (où on focalise sur la préservation de l'intégrité des données) et 1 (où l'objectif se résume à prévenir la reconstruction de S, peu importe le dommage).

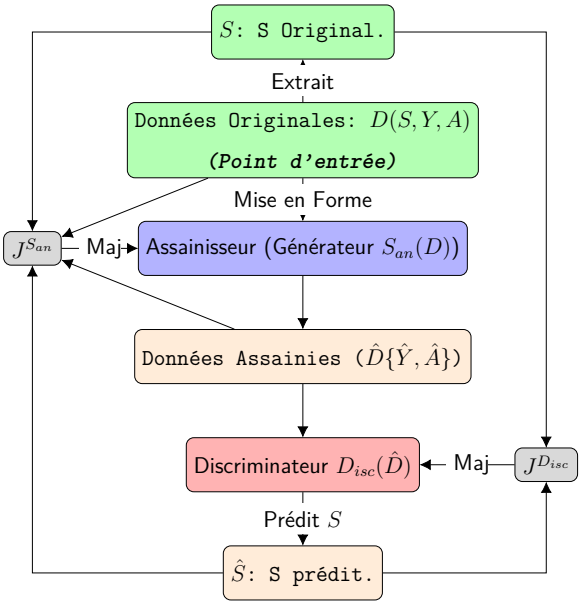


Figure 2.3 : Approche GANSan. L'objectif du discriminateur est de prédire S à partir des données $\hat{D}$ produites par l'assainisseur S_{an}. Les deux fonctions objectifs qui doivent être minimisées sont respectivement les fonction de coût du discriminateur et de l'assainisseur, $J^{D_{isc}}$ and $J^{S_{an}}$.

L'objectif du discriminateur peut s'écrire de la forme suivante :

$$J^{D_{isc}}(D,\ S_{an},\ D_{isc}) = d_{disc}(S,\ D_{isc}(S_{an}(D))) \tag{2.1}$$

En pratique nous avons utilisé la moyenne du carré des erreurs, ainsi que l'entropie des erreurs croisées comme choix de d_{disc}. En pratique, durant nos expérimentations, ces deux mesures d'erreur produisent des résultats similaires.

La fonction objectif de l'assainisseur est un peu plus complexe et est définie de la manière suivante :

$$J^{S_{an}}(D,\ S_{an},\ D_{isc}) = \alpha * d_s(S,\ D_{isc}(S_{an}(D))) + (1-\alpha) * d_r(D,\ S_{an}(D)) \tag{2.2}$$

dans laquelle :

— le coût de l'attribut protégé d_s permet de prendre en compte le retour du discriminateur. Dans notre approche d_s correspond soit au BER, soit à $sAcc$.

— d_r représente l'erreur de reconstruction des données, qui quantifie le dommage introduit par l'assainissement. La moyenne de la valeur absolue des erreurs (MAE) peut être utilisée pour instancier cette erreur.

L'assainisseur a été entraîné avec deux types d'optimisations dans nos expériences : une optimisation globale dans laquelle le réseau est mis à jour à partir de l'erreur totale calculée à partir de tous les attributs (la formulation précédente de $J^{S_{an}}$), et une optimisation locale basée sur l'utilisation des erreurs de chaque attribut, prises séparément. Pour un large choix de paramètres, nous avons observé que l'optimisation locale produit les meilleurs résultats, particulièrement lorsqu'elle est combinée avec l'optimiseur $Adam$ [Kingma et Ba, 2014].

La formulation de l'optimisation locale est la suivante :

$$\hat{D} = S_{an}(D) = \{\hat{A}, \hat{Y}\} \tag{2.3}$$

$$\hat{S} = D_{isc}(S_{an}(D)) \tag{2.4}$$

$$e_{A_i} = (1-\alpha) * |A_i - \hat{A}_i|; \quad \hat{A}_i \in \hat{A}, i \in [1,n] \tag{2.5}$$

$$e_Y = (1-\alpha) * |Y - \hat{Y}|; \tag{2.6}$$

$$e_S = \alpha * |S - \hat{S}|; \tag{2.7}$$

$$\vec{J}^{San} = (e_{A_1}, e_{A_2}, e_{A_3}, ..., e_{A_n}, e_Y, e_S)^T \tag{2.8}$$

$$\tag{2.9}$$

Cette procédure peut être vue comme une extension du cas type en apprentissage automatique où l'on essaie de prédire un attribut à partir plusieurs attributs.

Du point de vue de l'optimisation, l'erreur locale correspond à minimiser la fonction objectif suivante : $f(\{S, Y, A\}) = 1 - \frac{1}{N}\sum_N |S - \hat{S}|$ sous contraintes que $|a - \hat{a}| \leq \epsilon; a \in A, \hat{a} \in S_{an}(A)$, avec ϵ un seuil prédéfini.

2.4 *Description des ensembles de données*

Nous avons utilisé deux ensembles de données pour valider notre approche. Le premier dataset est *Adult Census Income*[1]. Cet ensemble de données comprend les informations collectées par les services de recensement aux États-Unis, concernant la situation financière d'individus. L'ensemble de données comprend 45222 profils caractérisés par 15 attributs tels que le genre, l'âge, le niveau d'éducation, l'origine ethnique et enfin le revenu. Pour notre analyse, l'attribut sensible considéré est le genre qui peut prendre les valeurs 0 ou 1 ($Homme$ ou $Femme$). La tâche de classification pour laquelle cet ensemble de données a été utilisé dans la communauté d'apprentissage automatique est celle de prédire si un individu aurait un revenu supérieur ou inférieur à $50K$ ($Y =' \leq 50K'$ ou $Y =' > 50K'$).

Le second ensemble de données est *German credit*[2], composé de 1000 individus et contenant des informations sur le profil bancaire de ces derniers, ainsi que sur d'autres attributs qui sont utiles pour évaluer la qualité d'un client sollicitant un crédit (par exemple, l'âge, l'historique de crédit, etc.). Au total, chaque individu est décrit par 21 attributs, notamment le type de client (*good* ou *bad*) que nous utiliserons en tant qu'attribut de décision dans notre approche. [Kamiran et Calders, 2009] ont trouvé que l'impact disparate DispImpact est maximale lorsque le jeu de données est partitionné entre jeunes et personnes plus âgées, le seuil est fixé à 25 ans. Le tableau 2.1 présente les distributions des différents groupes, subdivisés par rapport à S et par Y.

---

1. `https://archive.ics.uci.edu/ml/datasets/Statlog+\%28German+Credit+Data\%29`

2. `https://archive.ics.uci.edu/ml/datasets/adult`

Tableau 2.1 : Distribution des différents groupes relativement à l'attribut sensible et l'attribut de décision pour les ensembles Adult census et German credit.

Jeu de données	**Adult Census**		**German Credit**	
Groupe	Protégé ($S_x = S_0$)	Défaut ($S_x = S_1$)	Protégé ($S_x = S_0$)	Défaut ($S_x = S_1$)
$Pr(S = S_x)$	$36,21\%$	$63,79\%$	19%	81%
$Pr(S = S_x \mid Y = 1)$	$14,89\%$	$85,11\%$	$15,71\%$	84.29%
$Pr(Y = 1 \mid S = S_x)$	$11,35\%$	$31,24\%$	$57,89\%$	72.83%
$Pr(Y = 1, S = S_x)$	$3,69\%$	$21,09\%$	11%	59%
$Pr(Y = 1)$	$24,78\%$		70%	
$Pr(Y = 0)$	$75,22\%$		30%	

2.5 *Méthodologie d'évaluation*

Métriques d'équité. Pour évaluer l'équité, nous utiliserons 3 métriques, dont notamment le *Balanced Error Rate* (BER), le *Sensitive Accuracy* ($sAcc$), et l'impact disparate *Disparate Impact* ($DispImpact$). Ces métriques ont été précédemment décrites dans la section 1.2.

Mesures d'utilité. Pour quantifier le dommage introduit par l'assainissement, nous nous appuierons sur plusieurs mesures. La première de ces mesures est communément utilisée en apprentissage automatique, il sagit du taux de prédiction correcte d'un classifieur sur la prédiction de l'attribut de décision : $yAcc$. Cette métrique est définie de manière semblable au taux de prédictions justes de la prédiction de S ($sAcc$). La différence de cette métrique sur l'ensemble de données non modifié et celui assaini permet ainsi de quantifier la perte en utilité introduite par l'assainissement, l'objectif étant d'avoir un assainisseur préservant au mieux les informations nécessaires pour que Yreste utilisable. En plus de la perte en utilité, nous utiliserons aussi la distance entre

le jeu de données original et celui assaini, que nous appelons *distance d'assainissement*

$$san\_dist = \frac{1}{N \times d} \sum_{i=1}^{N} \sum_{j=1}^{d} (x_{i,j} - \hat{x}_{i,j})^2. \tag{2.10}$$

$x_{i,j}$ fait référence à la valeur du j^{ime} attribut du profil i, $\hat{x_{i,j}}$ correspond à la valeur assainie du j^{ime} attribut du profil i. Pour cette métrique, la base de référence est la distance entre le jeu de données original et celui obtenu par une génération aléatoire des valeurs.

Enfin, la dernière métrique d'utilité est la moyenne de la distance entre les individus dans le jeu assaini (*diversité*), que l'on compare avec la même métrique calculée sur le jeu de données original.

$$diversité = \frac{1}{(N-1) \times \sqrt{d}} \sum_{i=1}^{N} \sum_{j=1}^{N} \sqrt{\sum_{k=1}^{d} (\hat{x}_{i,k} - \hat{x}_{j,k})^2}. \tag{2.11}$$

Cette métrique permet de quantifier l'uniformité et la diversité des ensembles de données avant et après assainissement. Nous cherchons à vérifier que l'assainissement préserve le plus possible la diversité des valeurs des attributs du jeu original, car dans un cas extrême, l'assainisseur transforme tous les profils en un seul profil médian fixe, aboutissant en une perte complète en utilité ($diversité = 0$).

Mode opératoire. Pour évaluer la protection de S, nous avons entraîné quatre classifieurs différents : les *Machines à Vecteur de Support (SVM)* [Cortes et Vapnik, 1995], le *Perceptron Multicouches (MLP)* [Popescu *et al.*, 2009], *Gradient Boosting (GB)* [Friedman, 2002] et aussi le discriminateur interne de GANSan. Ces classifieurs, à l'exception du discriminateur, sont utilisés une fois l'assainissement terminé et sont entraînés sur les données ainsi assainies. Pour ces raisons, nous les étiquetterons en tant que *classifieurs externes*. Le même pipeline et les mêmes types de classifieurs externes sont utilisés pour le calcul des métriques d'utilité. La sélection du modèle

de l'assainisseur a été faite en variant le paramètre α, nous avons choisi l'époque qui produisit le meilleur BER et la meilleure utilité sur l'ensemble de validation.

Mise en forme des données. Cette étape de prétraitement consiste en l'application de procédures qui faciliteront la convergence des modèles, puisque nous utilisons des réseaux de neurones. Il s'agit notamment de la transformation des attributs catégoriques en attributs binaires (*one-hot encoding*), par exemple l'attribut *Éducation* dont le domaine est $Master, Doctorate$ est transformé en $Education_{Doctorate}$ et $Education_{Master}$ dont les domaines sont $\{0, 1\}$. Les attributs sont ensuite normalisés entre 0 et 1 ou -1 et 1 selon la fonction d'activation utilisée en couche de sortie du générateur. Sur le jeu de données Adult, nous avons appliqué un logarithme sur les attributs $capital - gain$ et $capital - loss$ avant leur normalisation. Cette étape est justifiée par le fait que ces attributs ont des distributions proches d'un Dirac delta [Dirac, 1981], avec les valeurs maximales de 9999 et 4356 respectivement, et des médianes de 0 (91% et 95% des profils ont des valeurs de 0 pour ces attributs). Le logarithme permet de réduire l'échelle des valeurs de ces attributs. Une fois l'assainissement terminé, l'opération inverse est effectuée, plus une étape de limitation des valeurs possibles pour faire correspondre les minima et maxima des attributs avant et après assainissement.

2.6 *Résultats*

Évaluation de l'équité. Le tableau 2.2 présente les résultats que nous avons obtenus avec les classifieurs externes pour les ensembles de données originaux et assainis. En observant les valeurs du BER, on constate qu'il est beaucoup plus difficile de prédire l'attribut protégé à partir du jeu assaini, l'assainissement semblant avoir doublé la protection de S, puisque les différents BER ont pratiquement doublé. Tous les résultats se rapprochent de la valeur optimale de protection $0,5$. Plus précisément, pour l'ensemble de données Adult census, la valeur minimale que l'on obtient indépendamment du clas-

sifieur est de $0,14$ sur le jeu original et $0,35$ sur le jeu assaini. Pareillement, sur German Credit, la valeur minimale est de $0,22$ sur le jeu original et de $0,39$ pour celui assaini. Quant au sAcc, on constate que les classifieurs ont des taux de prédiction correcte qui se rapprochent de plus en plus de la proportion de la classe majoritaire, voire sont en deçà. Les observations faites sur le BER s'appliquent aussi sur cette mesure. Ainsi, notre premier objectif qui est la protection de l'attribut sensible semble être atteint.

En observant la discrimination mesurée par DispImpact, on constate que GANSan semble réduire la discrimination pour les ensembles de données dans lesquels la discrimination était fortement présente. En effet, sur Adult où la discrimination minimale originale est de $0,17$, GANSan a produit une version assainie dans laquelle cette discrimination a été réduite jusqu'à la valeur de $0,11$. Le meilleur cas dans le jeu original ($DispImpact = 0,17$) est proche du pire cas dans le jeu assaini ($DispImpact = 0,19$). En revanche, sur German Credit, la discrimination originale minimale est de $0,018$ et le jeu assaini a une discrimination trois fois plus grande $0,051$. Ces résultats ne sont pas forcément surprenants si on se souvient que la prioriété d'équité que l'on cherche à assurer est la protection de l'attribut sensible et non l'équité de groupe.

Dans lensemble, l'analyse de l'équité montre que GANSan est en mesure d'assainir les données de façon à rendre plus difficile la prédiction de l'attribut sensible. Voyons maintenant ce qu'il en est du dommage associé.

Évaluation de l'utilité. Les métriques d'utilité sont présentées dans le tableau 2.3. La mesure *diversité* montre que la diversité des profils dans Adult (respectivement German Credit) a baissé de $3,5\%$ (respectivement $14,5\%$) suite à l'assainissement de l'assainissement. Nous pensons que la perte plus élevée causée dans German Crédit peut être justifiée (au moins partiellement) par sa petite taille (1000 profils), ce qui peut ne pas être suffisant pour permettre à GANSan de correctement apprendre la distribution des données.

Tableau 2.2 : Évaluation de l'équité produite par GANSan . *Ref.* correspond à la valeur de référence, tant dis que *Assaini* est la valeur que l'on obtient après assainissement.

Jeux de données	**classifieur**	**BER**		**sAcc**		**DispImpact**	
		Réf.	Assaini	Réf.	Assaini	Réf.	Assaini
Adult	GB	$0,1534$	$0,4005$	$0,8622$	$0,7033$	$0,1758$	$0,1155$
	SVM	$0,1420$	$0,3538$	$0,8270$	$0,6219$	$0,3774$	$0,1916$
	MLP	$0,1560$	$0,4160$	$0,8500$	$0,7024$	$0,1715$	$0,1291$
German Credit	GB	$0,3880$	$0,4726$	$0,8150$	$0,7750$	$0,0182$	$0,0519$
	SVM	$0,2247$	$0,3973$	$0,7400$	$0,5750$	$0,0657$	$0,0552$
	MLP	$0,3859$	$0,5013$	$0,8050$	$0,7300$	$0,0688$	$0,0752$

La distance d'assainissement *san_dist* nous montre que GANSan a un impact très faible sur la qualité des données (approximativement $0,015$ sur Adult et $0,0833$ sur German Credit). On constate que cette distance est largement plus petite comparée à la distance de référence.

Le taux de prédiction correcte dans la prédiction de la décision ($yAcc$) montre que l'assainissement entraîne avec lui une perte en utilité, de l'ordre de 1.17% sur Adult et 1.27% sur German Credit.

Analyse du dommage sur les attributs catégoriques et numériques. La figure 2.4 présente la fonction de répartition du pourcentage de profils modifiés par attribut catégorique. On observe notamment sur Adult Census que pour 80% de ces attributs, moins de 15% des profils sont modifiés. Sur German Credit, 80% des attributs catégoriques ont au plus 18% de profils modifiés. Pour chaque attribut numérique, le taux de dommage par profil est calculé comme la valeur absolue de la différence entre la valeur originale et celle assainie le tout ramené à l'amplitude de l'attribut (valeur maximale - valeur minimale). Ce taux de dommage est présenté en figure 2.5 pour

Tableau 2.3 : Évaluation de l'utilité de GANSan.

Dataset	**classifieurs**	**yAcc**		**san_dist**		*diversité*	
		Ref	Assaini	Ref	Assaini	Ref	Assaini
Adult	GB	$0,8692$	$0,8570$	$0,1363$	$0,0146$	$0,2905$	$0,2804$
	SVM	$0,7688$	$0,7848$				
	MLP	$0,8459$	$0,8590$				
German Credit	GB	$0,7900$	$0,7800$	$0,3314$	$0,0833$	$0,4906$	$0,4193$
	SVM	$0,6950$	$0,6700$				
	MLP	$0,7600$	$0,7500$				

l'ensemble de données Adult et 2.6 pour german credit. On y constate notamment que sur Adult, 88% des profils ont un taux de dommage inférieur à 13%. Ce taux est beaucoup plus important sur German Credit. En effet, pour certains attributs (par exemple, *Duration*, *InstallmentRate* et *ResidenceSince*), près de la moitié des profils ont un taux supérieur à 25%.

Comparaison avec l'état de l'art. Une des principales différences de GANSan par rapport aux autres approches est le fait que l'espace de représentation des données n'est pas changé, l'attribut sensible étant protégé tout en essayant de minimiser le dommage sur les données générées. Sur Adult (respectivement German Credit), on peut atteindre un taux de prédiction correcte de 85% (respectivement 78%) dans la prédiction de la décision. En comparaison, [Zemel *et al.*, 2013] et [Edwards et Storkey, 2015] peuvent avoir des taux de prédiction correcte respectifs de $0,78\%$ et $0,825\%$ sur Adult. Cependant, avec le changement d'espace, l'attribut sensible semble mieux protégé : GANSan obtient un $sAcc$ $0,7033$ au pire cas sur Adult, tandis que [Zemel *et al.*, 2013] obtiennent $0,55$.

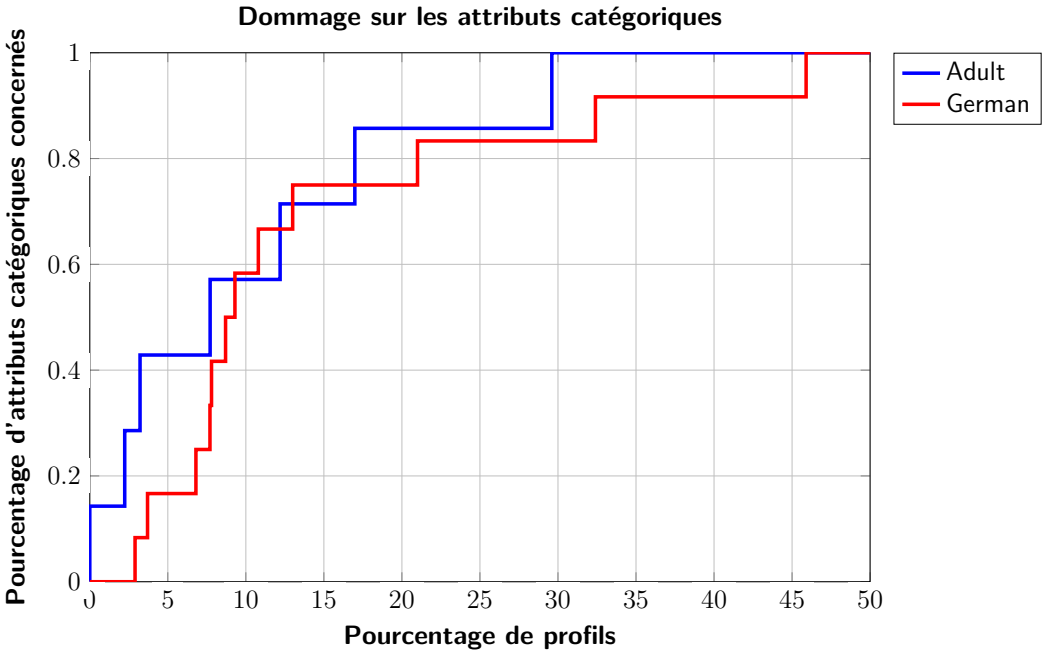


Figure 2.4 : Fonctions de répartition du pourcentage de profils modifiés par attribut catégorique.

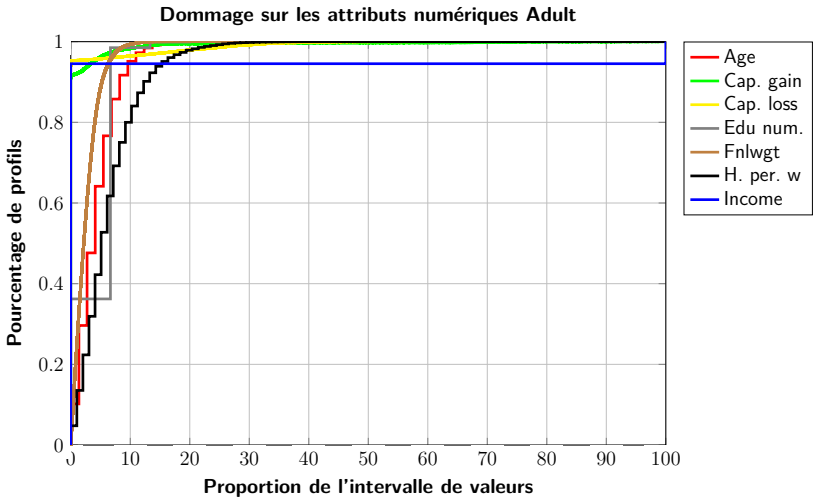


Figure 2.5 : Fonction de répartition du taux de dommage pour les attributs numériques sur Adult Census.

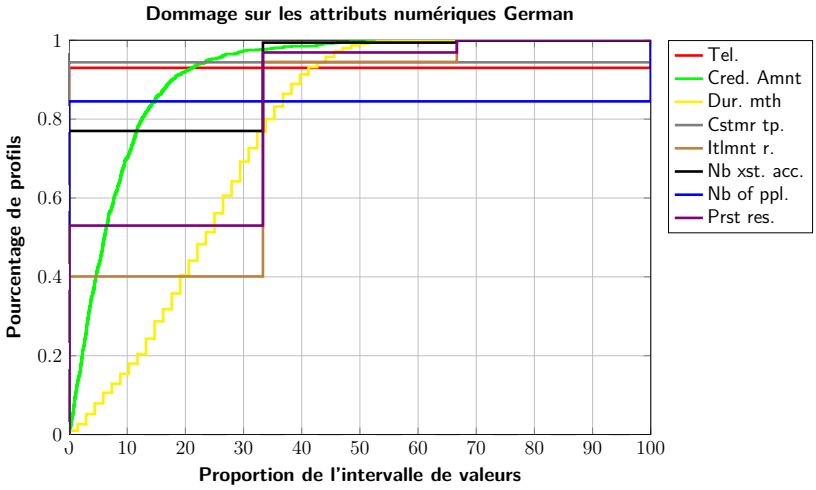


Figure 2.6 : Fonction de répartition du taux de dommage pour les attributs numériques sur German credit.

Les résultats que nous avons obtenus avec notre approche GANSan montrent que nous avons pu atteindre notre objectif qui est de protéger l'attribut sensible contre les inférences. Le dommage sur les données est assez faible, ce qui nous laisse penser que nous avons pu préserver l'utilité des données pour des tâches subséquentes. La comparaison avec l'état de l'art nous montre que nous ne sommes pas très éloignés des meilleurs résultats, bien que des améliorations soient encore nécessaires. Nous comptons soumettre notre approche *Assainissement des données en vue d'améliorer l'équité par des réseaux adversariaux génératifs* à la conférence Intelligence Artificielle, Éthique et Société (AIES [3])

Dans ce chapitre et le chapitre précédent, nous avons présenté des approches pour réduire la discrimination dans les ensembles de données. La *transparence*, qui est axe de recherche étroitement lié à la discrimination nécessite aussi d'être étudié. Cette proximité vient du fait que la *transparence* apporte des éléments de réponse sur le

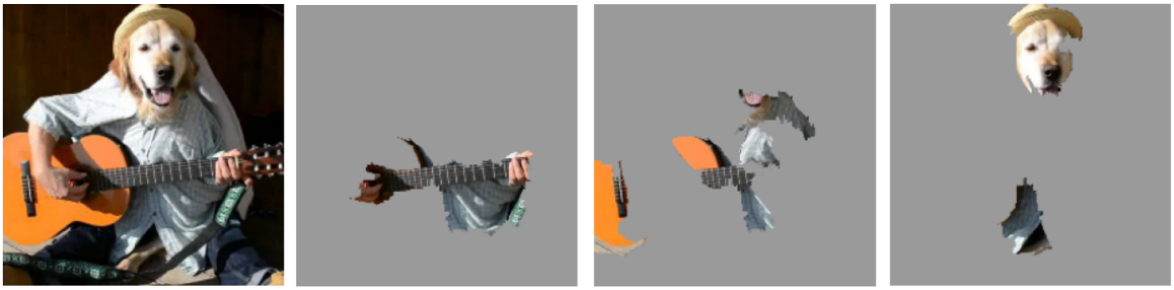
3. http://www.aies-conference.com/

fonctionnement des algorithmes, mais aussi sur la manière dont les données personnelles sont utilisées. Elle permettrait notamment de déceler si un processus utilise des attributs légalement prohibés dans un processus de décision en analysant le modèle utilisé et/ou ses entrées et sorties respectives. Dans le chapitre suivant, nous définirons ce que l'on entend par *transparence*, et nous présenterons quelques approches pour assurer l'interprétabilité des algorithmes.

CHAPITRE III

TRANSPARENCE DES ALGORITHMES

Le RGPD, en particulier les articles 15 - le droit d'accès à la donnée personnelle, qui offre en outre la possibilité à un sujet de questionner le système sur les décisions qui sont prises à son encontre, 17 - le droit à l'oubli et 18 - le droit à la restriction de traitement, en plus de la nécessité de confiance dans les décisions automatiques, renforce le besoin de transparence des algorithmes. Jenna Burrell [Burrell, 2016] relève notamment trois formes d'opacités : *l'opacité due à la protection du secret industriel*, *l'opacité due à l'analphabétisme technique*, et celle due à *la manière dont les algorithmes fonctionnent*. Une solution naïve serait la transparence par publication de code, cependant, celle-ci s'oppose directement aux deux premières formes d'opacités, et il peut arriver que le code reste toujours incompréhensible même pour des experts. Plusieurs approches ont été développées pour résoudre la troisième forme d'opacité. Nous pouvons citer par exemple les outils *LIME* et *SP-LIME* [Ribeiro *et al.*, 2016] pour expliquer une décision et pour expliquer le modèle boite noire respectivement. *LIME* effectue un échantillonnage autour de la donnée dont la décision doit être expliquée (pour une image par exemple, des super-pixels seront extraits), ensuite pour chacun des nouveaux échantillons, l'outil va effectuer une requête vers la boite noire pour obtenir une décision. Enfin, en utilisant la régression Lasso (qui permet de sélectionner uniquement certains coefficients tout en forçant les autres à avoir une valeur nulle), l'outil retiendra les échantillons qui ont les plus influencés la décision. De même, [Adler *et al.*, 2018] mesurent l'importance de



(a) Image Originale (b) Décision : Guitare Électrique (c) Décision : Guitare Acoustique (d) Décision : Labrador

Figure 3.1 : LIME, explications. Chacune des images (b), (c) et (d) illustre la combinaison de *superpixels* qui explique la décision.

chaque attribut en obfusquant l'attribut en question à l'aide d'une procédure similaire à celle de [Feldman *et al.*, 2015] (cf. section 1.3.3). Un attribut et ses corrélations ainsi masqués, on peut mesurer son importance sur la décision en mesurant la différence de taux de prédiction correcte de classification.

De manière plus générale, [Guidotti *et al.*, 2018] ont mené une étude sur les différentes méthodes pour expliquer les boites noires et relèvent notamment le fait que chacune de ces méthodes incorpore sa propre notion d'interprétabilité, et donc qu'il n'existe pas de consensus à l'heure actuelle. Ainsi, ils classifient les approches en fonction de la formulation du problème et du modèle d'explication utilisé. On distingue principalement quatre formulations :

— *l'explication du modèle boite noire*, qui consiste en l'approximation du modèle par un autre qui imite le comportement de la boite noire et qui est en soi compréhensible,

— *l'explicabilité des décisions du modèle*. Ici, il s'agit de trouver une explication pour une décision obtenue,

— *l'inspection de la boite noire*, où l'on cherche à trouver une représentation (visuelle ou textuelle) permettant de comprendre comment le modèle fonctionne et pourquoi certaines décisions sont plus plausibles que d'autres,

— *le design d'une boite noire transparente.* Il s'agit ici de construire un modèle intrinsèquement interprétable.

Les modèles d'explication utilisés sont très souvent des classifieurs qui ont déjà la propriété d'être interprétables. Par exemple, il s'agit des arbres de décisions, des règles (règles de décisions, listes de règles, ensembles de règles, etc.), mais on y retrouve aussi des prototypes qui résument des ensembles de points similaires, des graphiques de dépendance partielle (*PDP*), des cartographies de points fondamentaux (*saliency mask*), des mesures de l'importance des caractéristiques d'entrées, etc. On peut en plus raffiner la classification en fonction des types entrés (images, texte, données tabulaires, etc.) utilisés par ces méthodes. Un des critères importants est la disponibilité du modèle boite noire. En effet, il existe des situations où le modèle en boite noire n'est pas accessible, ou encore le fait de faire des requêtes successives déclenche des mécanismes de protection contre des attaques de propriété de sécurité (déni de service par exemple). La plupart des méthodes supposent que toutes les caractéristiques utilisées pour prendre une décision sont connues [Guidotti *et al.*, 2018]. Cependant, il existe des situations où la liste exhaustive des entrées n'est pas connue. D'autres données, en plus de celles qui sont présentées, sont utilisées en interne de la boite noire. La question de savoir ce qu'il est fait de nos données, et par qui sont-elles utilisées restent soulevées. Les systèmes tels que Sunlight [Lecuyer *et al.*, 2015], Xray [Lécuyer *et al.*, 2014], et AdFisher [Datta *et al.*, 2015] essaient d'apporter des éléments de réponses à ces questions. Les auteurs de Sunlight ont utilisé leur plateforme entre autres sur la messagerie de Google, Amazon et YouTube. Le système consiste à collecter des données en faisant varier plusieurs paramètres tels que l'historique de navigation, le contenu des messages électroniques, etc. Chaque variation donne lieu à un profil. Les sorties collectées étaient : les publicités ainsi que les produits recommandés. Ensuite, les attributs des profils sont ordonnés par rapport à leur influence sur les sorties en utilisant la pondération d'une régression Lasso. Enfin les explications sont générées par des combinaisons de conjonctions ou disjonctions et validées par des tests statistiques. Xray, qui est une version antérieure de

Sunlight fonctionne à peu près sur le même principe. Seulement, les hypothèses retenues sont celles qui couvrent un nombre significatif de profils.

Si l'on définit la transparence comme toute méthode permettant d'apporter des connaissances sur le fonctionnement interne de la boite noire, alors nous pouvons faire référence à des outils orientés vers la recherche de discrimination tout en gardant un haut niveau d'interprétabilité : FairTest [Tramèr *et al.*, 2015] et Themis [Galhotra *et al.*, 2017]. FairTest découpe dans un premier temps la population globale en sous-ensembles d'entraînement et de test. Dans le sous-ensemble d'entraînement, la partition de la population dans laquelle la corrélation est maximale avec l'attribut protégé est recherchée. Puis le découpage ayant permis de trouver ce sous-ensemble est validé statistiquement sur le jeu de test. Étant donné une vérité terrain, cet outil est aussi capable de trouver la population la plus sujette à des erreurs de classification. Cette dernière directive appelée *profilage d'erreur* permet notamment aux développeurs d'avoir des informations sur le processus de décision utilisé, en leur permettant d'analyser cette population à haut risque d'erreurs. Themis quant à lui cherche à établir la causalité des décisions obtenues. Pour cela, étant donné une liste d'attributs, l'outil va tester toutes les combinaisons possibles sans modifier ceux protégés, et mesurer ainsi la proportion de profils pour lesquels la décision obtenue a été différente. Les auteurs de Thémis ont notamment mis en évidence le fait que les approches qui réduisent la discrimination de groupe échouent au niveau du test de causalité. En d'autres termes, le fait d'utiliser une approche qui garantit l'équité de groupe n'empêche pas de trouver des profils qui ne diffèrent que sur une combinaison limitée d'attributs, mais pour lesquels les décisions prises sont différentes. Puisque le test de causalité mesure l'équité individuelle, nous pensons que ces échecs sont dus à l'incompatibilité entre cette équité et l'équité de groupe.

RÉFÉRENCES

AATIF, S. (2017). FaceApp : Selfie App accused of Racism for Whitening users' skin. Récupéré le 2017-04-25 de `http://www.independent.co.uk/life-style/gadgets-and-tech/news/faceapp-selfie-app-racism-filter-whitens-users-skin-viral-photo-a7701036.html`

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B. et Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, *54*(1), 95–122.

Angwin, J., Larson, J., Mattu, S. et Kirchner, L. (2016). Machine Bias. Récupéré le 2016-05-23 de `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`

Baldi, P. et Hornik, K. (1989). Neural networks and principal component analysis : Learning from examples without local minima. *Neural networks*, *2*(1), 53–58.

Barocas, S. et Selbst, A. D. (2016). Big data's disparate impact. *Cal. L. Rev.*, *104*, 671.

Bilal Zafar, M., Valera, I., Gomez Rodriguez, M. et Gummadi, K. P. (2017). Fairness Constraints : Mechanisms for Fair Classification.

Burrell, J. (2016). How the machine thinks : Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Calders, T., Karim, A., Kamiran, F., Ali, W. et Zhang, X. (2013). Controlling attribute effect in linear regression. Dans *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 71–80. IEEE.

Carrascosa, J. M., Mikians, J., Cuevas, R., Erramilli, V. et Laoutaris, N. (2015). I always feel like somebody's watching me : measuring online behavioural advertising. Dans *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, p. 13. ACM.

Center, E. P. I. (2016). EPIC - Algorithms in the Criminal Justice System. Récupéré le 2017-04-17 de `https://epic.org/algorithmic-transparency/crim-justice/`

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Datta, A., Tschantz, M. C. et Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, *2015*(1), 92–112.

Dave Gershgorn (2017). AI is now so complex its creators cant trust why it makes decisions. Récupéré le 2017-12-07 de `https://qz.com/1146753/ai-is-now-so-complex-its-creators-cant-trust-why-it-makes-decisions/`

Davide Castelvecchi (2016). Can we open the black box of AI ? Récupéré le 2016-10-05 de `https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731`

Dirac, P. A. M. (1981). *The principles of quantum mechanics*. Numéro 27. Oxford university press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. et Zemel, R. (2012). Fairness through awareness. Dans *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

Edwards, H. et Storkey, A. (2015). Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Englehardt, S. et Narayanan, A. (2016). Online tracking : A 1-million-site measurement and analysis. Dans *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, 1388–1401., New York, NY, USA. ACM. `http://dx.doi.org/10.1145/2976749.2978313`. Récupéré de `http://doi.acm.org/10.1145/2976749.2978313`

Faliagka, E., Tsakalidis, A. et Tzimas, G. (2012). An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet research*, *22*(5), 551–568.

Fathy, M. E., Patel, V. M. et Chellappa, R. (2015). Face-based active authentication on mobile devices. Dans *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 1687–1691. IEEE.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. et Venkatasubramanian, S. (2015). Certifying and removing disparate impact. Dans *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P. et Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning. *ArXiv e-prints*.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Galhotra, S., Brun, Y. et Meliou, A. (2017). Fairness testing : testing software for discrimination. Dans *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 498–510. ACM.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. et Bengio, Y. (2014). Generative adversarial nets. Dans *Advances in neural information processing systems*, 2672–2680.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. et Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), 93.

Guynn, J. (2015). Google Photos labeled black people 'gorillas'. Récupéré le 2017-07-17 de `http://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/`

Hardt, M., Price, E., Srebro, N. *et al.* (2016). Equality of opportunity in supervised learning. Dans *Advances in neural information processing systems*, 3315–3323.

Hindupur, A. (2018). the-gan-zoo. Récupéré le 2018-08-11 de `https://github.com/hindupuravinash/the-gan-zoo/blob/master/gans.tsv`

Ingold, D. et Soper, S. (2016). Amazon Doesnt Consider the Race of Its Customers. Should It ? Récupéré le 2016-04-21 de `https://www.bloomberg.com/graphics/2016-amazon-same-day/`

Kamiran, F. et Calders, T. (2009). Classifying without discriminating. Dans *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, 1–6. IEEE.

Kamiran, F. et Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33.

Kamiran, F., Calders, T. et Pechenizkiy, M. (2010). Discrimination aware decision tree learning. Dans *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, 869–874., Washington, DC, USA. IEEE Computer Society. `http://dx.doi.org/10.1109/ICDM.2010.50`. Récupéré de `http://dx.doi.org/10.1109/ICDM.2010.50`

Kamishima, T., Akaho, S., Asoh, H. et Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. Dans *Joint European Conference on Machine*

Learning and Knowledge Discovery in Databases, 35–50. Springer.

Kingma, D. P. et Ba, J. (2014). Adam : A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. et Welling, M. (2013). Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Laperdrix, P., Rudametkin, W. et Baudry, B. (2016). Beauty and the beast : Diverting modern web browsers to build unique browser fingerprints. Dans *Security and Privacy (SP), 2016 IEEE Symposium on*, 878–894. IEEE.

Lécuyer, M., Ducoffe, G., Lan, F., Papancea, A., Petsios, T., Spahn, R., Chaintreau, A. et Geambasu, R. (2014). Xray : Enhancing the web's transparency with differential correlation. Dans *USENIX Security Symposium*, 49–64.

Lecuyer, M., Spahn, R., Spiliopolous, Y., Chaintreau, A., Geambasu, R. et Hsu, D. (2015). Sunlight : Fine-grained targeting detection at scale with statistical confidence. Dans *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 554–566. ACM.

Louizos, C., Swersky, K., Li, Y., Welling, M. et Zemel, R. (2015). The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Luong, B. T., Ruggieri, S. et Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. Dans *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 502–510. ACM.

Mahmoud, M., Algadi, N. et Ali, A. (2008). Expert system for banking credit decision.

Malinas, G. et Bigelow, J. (2004). Simpson's paradox.

Martello, S. (1990). Knapsack problems : algorithms and computer implementations. *Wiley-Interscience series in discrete mathematics and optimiza tion*.

Mirza, M. et Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Ninghui, L., Tiancheng, L. et Venkatasubramanian, S. (2007). *t-Closeness : Privacy beyond k-anonymity and l-diversity*, Dans *Proceedings - International Conference on Data Engineering*, (p. 106–115).

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L. et Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, *8*(7), 579–588.

Ribeiro, M. T., Singh, S. et Guestrin, C. (2016). Why should i trust you ?: Explaining the predictions of any classifier. Dans *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. ACM.

Ruggieri, S. (2014). Using t-closeness anonymity to control for non-discrimination. *Trans. Data Privacy*, *7*(2), 99–129.

Tramèr, F., Atlidakis, V., Geambasu, R., Hsu, D. J., Hubaux, J.-P., Humbert, M., Juels, A. et Lin, H. (2015). Discovering unwarranted associations in data-driven applications with the fairtest testing toolkit. *CoRR, abs/1510.02377*.

Verma, S. et Rubin, J. (2018). Fairness definitions explained.

Wallach, H. (2014). Big Data, Machine Learning, and the Social Sciences. Récupéré le 2017-04-17 de `https://medium.com/@hannawallach/big-data-machine-learning-and-the-social-sciences-927a8e20460d`

Yu, L., Wang, S. et Lai, K. K. (2007). Basic learning principles of artificial neural networks. *Foreign-exchange-rate forecasting with artificial neural networks*, 27–37.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T. et Dwork, C. (2013). Learning fair representations. Dans *International Conference on Machine Learning*, 325–333.

Zhang, L., Wu, Y. et Wu, X. (2016). Situation testing-based discrimination discovery : A causal inference approach. Dans *IJCAI*, volume 16, 2718–2724.

Zhu, J.-Y., Park, T., Isola, P. et Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, *abs/1703.10593*.