Exploiter l'équité d'un modèle d'apprentissage pour reconstruire les attributs sensibles de son ensemble d'entraînement

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, Mohamed Siala

Exploiter l'équité d'un modèle d'apprentissage pour reconstruire les attributs sensibles de son ensemble d'entraînement

J. Ferry[1], U. Aïvodji[2], S. Gambs[3], M-J. Huguet[1], M. Siala[1]

[1] LAAS-CNRS, Université de Toulouse, CNRS, INSA, Toulouse, France
[2] École de Technologie Supérieure, Montréal, Canada
[3] Université du Québec à Montréal, Montréal, Canada

jferry@laas.fr

Résumé

Pour palier les biais non désirés en apprentissage supervisé, de nombreux travaux utilisent des métriques d'équité statistique, définies vis-à-vis de certains attributs sensibles. *Bien que ceux-ci ne soient généralement pas utilisés par le modèle appris au moment de l'inférence, ils le sont souvent pendant son entraînement pour contrôler l'équité. Nous montrons ainsi qu'un attaquant disposant d'un accès en boîte noire à un tel modèle peut utiliser le fait qu'il soit équitable pour reconstruire les attributs sensibles de son ensemble d'entraînement. L'approche proposée consiste à* corriger *une première reconstruction effectuée par un attaquant de la littérature, pour se conformer avec l'information de l'équité. Notre large évaluation expérimentale confirme que ce processus de correction permet d'améliorer les performances de l'attaque de manière significative.*

Mots-clés

Attaque de reconstruction, vie privée, équité, apprentissage, programmation linéaire en nombres entiers, programmation par contraintes.

Abstract

To face the undesirable biases in machine learning, a growing body of work consider statistical fairness metrics, defined with respect to some sensitive *attributes. Even though the later are generally not used by the learnt model for inference, they are often required at training time to ensure fairness. We then show that an adversary with black-box access to such model can leverage the fact that it is fair to reconstruct the sensitive attributes of its training set. The proposed approach consists in* correcting *a baseline reconstruction made by some adversary from the literature to comply with the fairness information. Our thorough experimental study demonstrates that this correction process significantly improves the performances of the performed attack.*

Keywords

Reconstruction attack, privacy, fairness, machine learning, integer linear programming, constraint programming.

1 Introduction

L'utilisation croissante de modèles d'apprentissage pour la prise de décisions à forts enjeux (*e.g.,* admissions à l'université, prédiction de récidive…) soulève de nombreux risques éthiques, parmi lesquels celui de discrimination. Pour faire face à cette problématique, de nombreux travaux proposent d'apprendre des modèles respectant des contraintes d'équité, exprimées vis-à-vis de certains attributs *sensibles* [3, 7, 23]. Ces derniers correspondent aux caractéristiques telles que le genre, l'âge ou l'origine ethnique [10], qui ne devraient pas influer sur les processus de prise de décision impactant les individus, pour des raisons légales, éthiques, sociales ou philosophiques [3].

Un autre aspect fondamental pour un apprentissage responsable est la protection de la vie privée. En effet, les modèles d'apprentissage sont souvent entraînés sur de grandes quantités de données personnelles. Il est alors important de s'assurer que ces modèles apprennent des motifs génériques utiles, sans révéler d'informations privées sur un ou plusieurs individus en particulier. Dans ce contexte, les *attaques d'inférence* [9, 14] visent à utiliser le résultat d'un calcul (*e.g.,* un modèle entraîné) pour retrouver des informations sur ses entrées (*e.g.,* un ensemble d'entraînement). Notre travail appartient à la famille des *attaques de reconstruction* (de jeux de données), dans lesquelles un attaquant essaie de reconstruire tout ou partie de l'ensemble d'entraînement d'un modèle [9]. Nous considérons le cas dans lequel l'attaquant tente de reconstruire la colonne des attributs sensibles de l'ensemble d'entraînement.

En fonction des données en sa possession *(connaissances adversariales)*, un attaquant peut adopter différentes stratégies pour reconstruire les attributs sensibles de l'ensemble d'entraînement d'un modèle. Nous proposons une méthode de post-traitement intitulée *correction de reconstruction*, qui prend en entrée une reconstruction initiale effectuée par un attaquant de la littérature, éventuellement associée à des scores de confiance pour chaque élément. Notre méthode modifie ensuite cette reconstruction initiale de manière à se conformer avec certaines contraintes définies par l'utilisateur. Notre travail considère le scénario dans lequel ces contraintes sont des contraintes d'équité, et l'attaquant uti-

TABLE 1 – Résumé des métriques d'équité considérées

Ref.	Métrique	Mesure égalisée	Expression de la contrainte
[13]	Statistical Parity (SP)	Probabilité de prédiction positive	$\forall s,\ \lvert\mathbb{P}(\hat{y}=1) - \mathbb{P}(\hat{y}=1 \mid s)\rvert \leq \epsilon$
[8]	Predictive Equality (PE)	Taux de Faux Positifs	$\forall s,\ \lvert\mathbb{P}(\hat{y}=1 \mid y=0) - \mathbb{P}(\hat{y}=1 \mid s,\ y=0)\rvert \leq \epsilon$
[19]	Equal Opportunity (EO)	Taux de Vrais Positifs	$\forall s,\ \lvert\mathbb{P}(\hat{y}=1 \mid y=1) - \mathbb{P}(\hat{y}=1 \mid s,\ y=1)\rvert \leq \epsilon$
[19]	Equalized Odds (EOdds)	Taux de Faux Positifs et de Vrais Positifs	Conjonction de la Predictive Equality et de l'Equal Opportunity

lise le fait qu'un modèle soit équitable pour améliorer sa reconstruction initiale. Cette *information de l'équité* peut être liée à des obligations légales, comme par exemple la "règle des 80 pourcents" [16] de la Commission américaine pour l'égalité des chances dans l'emploi *(US Equal Employment Opportunity Commission)* [15].

Les tensions entre équité et protection de la vie privée en apprentissage ont été récemment étudiées à travers les conflits théoriques et pratiques entre les métriques d'équité statistique et la confidentialité différentielle, méthode standard en protection de la vie privée [18]. Notre travail prend une direction différente mais démontre également que le fait d'imposer des contraintes d'équité pendant l'apprentissage peut compromettre la confidentialité des attributs sensibles. Ce travail complet a été présenté à la conférence internationale SATML 2023 *(The First IEEE Conference on Secure and Trustworthy Machine Learning)* [17].

2 Contexte et travaux antérieurs

2.1 Classification équitable

Soit M le nombre d'*attributs non sensibles* caractérisant un exemple. Pour $j \in \{1..M\}$, $\mathcal{X}_j$ est le domaine des valeurs possibles pour l'attribut j, qui peut être catégorique ou numérique, et $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_M$. De même, soit $\mathcal{S}$ (*resp.,* $\mathcal{Y}$) le domaine d'un *attribut sensible* (catégorique) (*resp., label*). $D = (X, S, Y)$ est un jeu de données issu d'une distribution sous-jacente (inconnue) sur $\mathcal{X} \times \mathcal{S} \times \mathcal{Y}$. Soit N le nombre d'*exemples* dans D, où $e_{i \in \{1..N\}} = (x, s, y) \in \mathcal{X} \times \mathcal{S} \times \mathcal{Y}$.

L'objectif d'un algorithme d'apprentissage supervisé est de construire un *classifieur* $\mathcal{L}(D) = h$ mappant l'espace des attributs à celui des labels. L'utilisation explicite d'un attribut sensible est en général interdite par la loi pour éviter un *traitement disparate* [4]. On considère donc que l'attribut sensible (utilisé pendant l'entraînement pour s'assurer de l'équité du modèle construit) n'est pas utilisé pour l'inférence : $h : \mathcal{X} \mapsto \mathcal{Y}$, et $\hat{Y} = h(X)$ sont les prédictions du modèle. De manière consistante avec la littérature de l'équité en apprentissage, on considère la tâche de classification binaire dans ce travail : $\mathcal{Y} = \{0, 1\}$. Néanmoins, notre approche peut facilement être étendue au cas de la classification non-binaire, étant donné des contraintes d'équité formulées dans ce cadre plus général.

Pour s'assurer que les modèles d'apprentissage ne reproduisent ni ne créent de *biais indésirables* (*e.g.*, menant à des discriminations), différentes notions d'équité ont été proposées dans la littérature [23]. Dans ce travail, nous considérons le cas dans lequel l'équité est exprimée avec des *métriques d'équité statistique* [13]. Celles-ci visent à s'assurer qu'une grandeur (*e.g.*, taux de vrais positifs) diffère d'au plus une valeur de tolérance donnée ϵ entre différents *groupes protégés* (définis par les attributs sensibles). De nombreuses méthodes ont été proposées [3, 7, 23] afin d'apprendre des modèles *équitables*. Elles peuvent être divisées en trois principales catégories, en fonction de l'étape du pipeline d'apprentissage à laquelle elles interviennent [5]. Les méthodes de *pre-processing* visent à atténuer les corrélations indésirables directement dans le jeu de données d'entraînement, avant d'utiliser des techniques classiques d'apprentissage sur le jeu de données ainsi modifié [21]. Les techniques d'*in-processing* consistent à modifier l'algorithme d'apprentissage lui-même, afin d'apprendre des modèles équitables. Enfin, les approches de *post-processing* [19] modifient les sorties d'un modèle préentraîné pour satisfaire certains critères d'équité.

Notre approche est agnostique au type de méthode utilisée pour apprendre le modèle équitable, puisque qu'elle dépend uniquement de ses prédictions et de l'information sur son équité. Nous utilisons quatre métriques d'équité statistique populaires dans la littérature : la Statistical Parity [13], la Predictive Equality [8], l'Equal Opportunity [19] et l'Equalized Odds [19]. Ces métriques sont présentées dans la Table 1, ainsi que les mesures qu'elles visent à égaliser entre les différents groupes protégés, et l'expression mathématique des contraintes associées.

2.2 Attaques de reconstruction

Notre approche est une *attaque d'inférence* [14], visant à retrouver des informations sur un jeu de données en observant les sorties d'un calcul sur celui-ci. Ici, le calcul en question est un algorithme d'apprentissage, et sa sortie est un modèle entraîné. Différents types d'attaques d'inférence ont été proposés contre les modèles d'apprentissage [9]. Notre attaque d'inférence est une *attaque de reconstruction* [9]. Elle nécessite seulement un accès en *boîte noire* au modèle équitable entraîné (*i.e.,* via une API de prédiction) et est agnostique au type de modèle, à l'algorithme d'apprentissage et à la méthode utilisée pour assurer l'équité.

Les attaques de reconstruction ont d'abord été étudiées dans le contexte des mécanismes d'accès aux bases de données. Dans le scénario considéré, une base de données contient des informations sur des individus, chaque exemple étant composé d'informations non privées ainsi que d'un *bit privé* (un par exemple/individu) [14]. L'objectif d'un attaquant est alors de reconstruire la colonne des *bits privés* de la base de données. Pour cela, il effectue des requêtes au mécanisme d'accès, dont les sorties sont des agrégations bruitées des bits privés des individus. De telles attaques de reconstruction ont été introduites et formalisées dans [11],

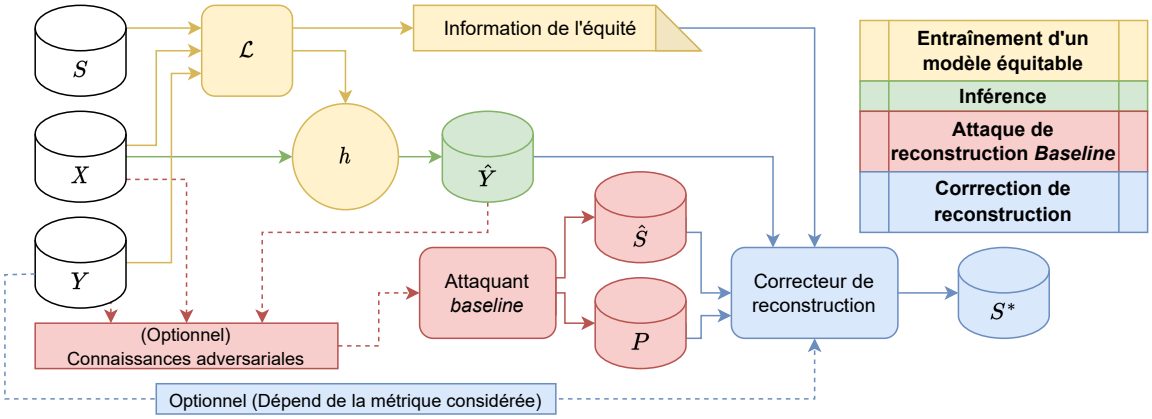


FIGURE 1 – Schéma de l'attaque proposée. Un modèle h est construit par la procédure d'apprentissage équitable $\mathcal{L}$ et utilisé pour l'inférence. Ensuite, un attaquant *baseline* tente de reconstruire les attributs sensibles S de l'ensemble d'entraînement de h. Cet attaquant produit une reconstruction $\hat{S}$ accompagnée de scores de confiance (optionnels) P. Notre contribution se situe au niveau du *Correcteur de reconstruction*, qui prend en entrée la reconstruction de l'attaquant *baseline* $\hat{S}$ et la modifie pour se conformer à l'information de l'équité, produisant ainsi S^*, la version corrigée de la reconstruction des attributs sensibles.

où elles sont formulées en utilisant un programme linéaire. Notre objectif est similaire à ces travaux : nous voulons reconstruire une *colonne* du jeu de données en utilisant la sortie d'un calcul utilisant cette colonne. Cependant, une différence fondamentale se trouve dans la nature du mécanisme accédant les données privées. D'un côté, les mécanismes d'accès aux bases de données considérés utilisent l'information privée pour calculer le résultat de chaque requête. De l'autre, dans notre cas, les attributs sensibles ne sont plus jamais utilisés au moment de l'inférence, et toute l'information les concernant est dévoilée en une seule fois, à travers le modèle entraîné lui-même ou ses prédictions.

Deux travaux sont proches du scénario considéré [1, 20]. D'un côté, [1] propose une attaque pour inférer les attributs sensibles d'un exemple étant donné les sorties d'un modèle pour cet exemple. En résumé, l'attaquant entraîne un modèle d'apprentissage en utilisant un *ensemble d'attaque* séparé, pour lequel il connaît les attributs sensibles. Cette attaque correspond à l'attaquant *baseline* considéré dans nos expériences (*cf.* Section 4.1). De l'autre côté, [20] propose un mécanisme dont le principe est assez proche de notre travail, mais considère une configuration très particulière où l'apprentissage est effectué de manière distribuée entre deux entités : un *learner* souhaitant apprendre un modèle équitable sur un certain jeu de données pour lequel il ne connaît pas les attributs sensibles, et un tiers qui les connaît. Le *learner* envoie itérativement les paramètres du modèle qu'il est en train de construire à ce tiers, qui lui indique si le modèle actuel respecte les contraintes d'équité. Il encode ensuite cette séquence d'informations d'équité dans un modèle de programmation en nombres entiers pour effectuer la reconstruction des attributs sensibles de l'ensemble d'entraînement. Tandis que l'intuition est similaire, notre travail couvre un cas d'usage plus général (en ne faisant aucune hypothèse sur l'algorithme d'apprentissage équitable utilisé) dans un cadre moins favorable (l'attaquant possédant uniquement l'information d'équité sur le modèle final).

3 Améliorer une reconstruction des attributs sensibles grâce à l'équité

Nous décrivons à présent l'approche proposée et montrons comment l'information sur l'équité d'un modèle peut être utilisée pour améliorer une reconstruction des attributs sensibles de son ensemble d'entraînement. Nous présentons le cadre considéré et positionnons le composant de *correction de reconstruction* que nous proposons, avant de décrire deux implémentations possibles pour celui-ci.

3.1 Principe de l'attaque

La Figure 1 illustre les différents composants de l'approche considérée. A partir d'un ensemble $D = (X, S, Y)$, un modèle h est entraîné par un algorithme d'apprentissage équitable $\mathcal{L}$, qui s'assure que h est équitable sur D selon une métrique d'équité définie par rapport à un certain attribut sensible S. Le classifieur h n'utilise pas l'attribut sensible S pour l'inférence afin d'éviter le *traitement disparate* [4]. Ainsi, h effectue ses prédictions en utilisant uniquement les attributs non sensibles X. Notre approche ne fait aucune hypothèse sur l'algorithme d'apprentissage équitable utilisé $\mathcal{L}$. En effet, le seul pré-requis de notre attaque est la connaissance de l'information de l'équité.

L'objectif de l'attaque est de reconstruire les attributs sensibles S de l'ensemble d'entraînement. Dans le schéma considéré, S est seulement utilisé par $\mathcal{L}$ pour s'assurer de l'équité de h (et n'est plus utilisé par la suite). Dans une première étape de l'attaque, un *attaquant baseline* effectue une première reconstruction $\hat{S}$ de S, en utilisant éventuellement certaines connaissances auxiliaires. L'attaquant *baseline* produit également un vecteur de probabilités P, qui reflète sa confiance dans chaque composant de sa reconstruction $\hat{S}$. Si l'attaquant ne calcule pas de scores de confiance, alors P correspond simplement au vecteur identité. Dans une seconde étape de l'attaque, un *correcteur de reconstruction* prend en entrée la reconstruction de l'at-

taquant *baseline* $\hat{S}$ et ses scores de confiance P. Il produit une nouvelle reconstruction S^* minimisant les changements (pondérés par les scores de confiance) par rapport à la reconstruction de l'attaquant *baseline*, tout en s'assurant du respect de certaines propriétés, telles que des contraintes d'équité statistique. Pour s'assurer du respect de telles contraintes, le *correcteur de reconstruction* nécessite également l'information de l'équité, les prédictions du modèle cible $\hat{Y}$ sur son ensemble d'entraînement, ainsi que (selon la métrique d'équité considérée, *cf.* Table 1) les vrais labels Y. Aucune hypothèse n'est faite sur le modèle cible h, qui peut être vu comme une boîte noire puisque l'attaque nécessite seulement l'accès à ses prédictions.

Le succès de l'attaque peut être évalué comme la *précision de la reconstruction* de S^* (*i.e.*, proportion d'éléments de S correctement prédits dans S^*). La contribution de notre travail se situe au niveau du *correcteur de reconstruction* qui, en incorporant uniquement l'information de l'équité, est capable d'améliorer significativement la qualité de la reconstruction des attributs sensibles. Cette amélioration peut être quantifiée en comparant la précision de la reconstruction de l'attaquant *baseline* $\hat{S}$ et celle de la version corrigée S^*.

3.2 Correcteur de reconstruction général

Nous présentons maintenant $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$, un modèle de Programmation Linéaire en Nombres Entiers implémentant le correcteur de reconstruction de la Figure 1, dans le cas d'un attribut sensible binaire. Son objectif est de modifier la reconstruction des attributs sensibles de l'ensemble d'entraînement de l'attaquant *baseline* pour se conformer à certaines contraintes (ici, l'information de l'équité) tout en minimisant les changements effectués (pondérés par la confiance de l'attaquant).

Entrées

- $\hat{s}_i \in \{0,1\},\ i = 1, \ldots, N$ (reconstruction initiale de l'attaquant *baseline*)
- $p_i \in \{0,1\},\ i = 1, \ldots, N$ (confiance de l'attaquant *baseline* pour $\hat{s}_i$)
- $\hat{y}_i \in \{0,1\},\ i = 1, \ldots, N$ (prédictions de h)
- Information de l'équité : h respecte une contrainte d'équité selon une métrique (*e.g.*, SP) et une valeur de tolérance ϵ

Variables de décision

- $s_i^* \in \{0,1\},\ i = 1, \ldots, N$ (reconstruction corrigée)

Modèle $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$

$$\min \sum_{i=1}^{N} (p_i \cdot (1 - \hat{s}_i) \cdot s_i^*) + \sum_{i=1}^{N} (p_i \cdot \hat{s}_i \cdot (1 - s_i^*)) \quad (1)$$

$$s.t. : 0 < \sum_{i=1}^{N} s_i^* < N \quad (2)$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{\sum_{i=1}^{N} \hat{y}_i \cdot s_i^*}{\sum_{i=1}^{N} s_i^*} \leq \epsilon \quad (3)$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{\sum_{i=1}^{N} \hat{y}_i \cdot (1 - s_i^*)}{\sum_{i=1}^{N} (1 - s_i^*)} \leq \epsilon \quad (4)$$

L'objectif (1) vise à minimiser les changements à $\hat{S}$ (pondérés par la confiance de l'attaquant). Chaque modification d'un élément $\hat{s}_i$ de la reconstruction initiale de l'attaquant *baseline* est pénalisée avec un coût p_i et le modèle minimise le coût total. La contrainte (2) s'assure que la reconstruction contienne au moins un exemple dans chaque groupe protégé. Enfin, les contraintes (3) et (4) encodent la métrique d'équité Statistical Parity. La constrainte (3) (*resp.,* la contrainte (4)) s'assure que le Taux de Prédictions Positives (TPP) sur le groupe 1 (*resp.,*, sur le groupe 0) diffère d'au plus ϵ du TPP global. Les prédictions $\hat{y}_i$ du modèle étant fixées, la contrainte d'équité est satisfaite en modifiant la reconstruction des attributs sensibles s_i^*.

Finalement, une solution optimale à notre modèle général de correction de reconstruction $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ est une affectation des variables binaires s_i^* minimisant (1) tout en satisfaisant les contraintes (2) à (4). L'affectation S^* correspond aux changements de coût minimum à la reconstruction initiale de l'attaquant *baseline* $\hat{S}$ de sorte à satisfaire les contraintes d'équité. Si les changements effectués sont corrects la majorité du temps (ce qui est attendu si les scores de confiance fournis par l'attaquant *baseline* sont de qualité), alors la précision globale de la reconstruction sera améliorée. Dans tous les cas, l'algorithme est garanti de trouver une solution satisfaisant les contraintes d'équité - ce qui n'était pas forcément le cas de la reconstruction *baseline*. Par ailleurs, puisqu'il est capable de modifier les attributs sensibles de tous les exemples d'entraînement, le modèle peut atteindre n'importe quelle valeur des métriques d'équité dans la version corrigée de la reconstruction. Ainsi, la connaissance précise de la violation de l'équité (plutôt qu'une simple borne supérieure via ϵ) pourrait être utilisée pour réduire encore l'espace des reconstructions admissibles et améliorer les performances de la correction de reconstruction. Enfin, puisqu'il encode explicitement l'attribut sensible de chaque exemple de l'ensemble d'entraînement, $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ peut être utilisé pour formuler n'importe quelle contrainte sur ces derniers.

3.3 Correcteur de reconstruction efficace

L'espace de recherche du modèle général de correction de reconstruction $\mathcal{RC}(\hat{S}, P, \hat{Y}, \epsilon)$ grandit exponentiellement avec le nombre d'exemples d'entraînement N. En effet, chaque élément du vecteur des attributs sensibles S étant considéré indépendamment des autres (et représenté comme une variable de décision binaire), la taille de l'espace de recherche est en $O(2^N)$, ce qui limite le passage à l'échelle. Cependant, une telle granularité n'est pas nécessaire lorsqu'on considère des métriques d'équité statistique. Plus précisément, pour satisfaire la contrainte d'équité, le correcteur de reconstruction peut réaliser exactement quatre actions différentes : (i) changer un élément de la reconstruction initiale $\hat{s}_i$ de 1 en 0 pour un exemple tel que $\hat{y}_i = 1$ (ii) changer $\hat{s}_i$ de 0 en 1, pour un exemple tel que $\hat{y}_i = 1$, (iii) changer $\hat{s}_i$ de 1 en 0, pour un exemple tel que $\hat{y}_i = 0$, ou (iv) changer $\hat{s}_i$ de 0 en 1, pour un exemple tel que $\hat{y}_i = 0$. Par ailleurs, pour l'opération choisie, le modèle commencera toujours par les exemples avec les scores de confiance les

plus faibles (puisqu'on minimise le coût des changements). Soit n_1^+ le nombre d'exemples d'entraînement prédits positivement par le modèle cible h et assignés au groupe 1 dans la reconstruction initiale de l'attaquant : $n_1^+ = \sum_{i=1}^{N} \hat{s}_i \cdot \hat{y}_i$. De même, soit $n_0^+ = \sum_{i=1}^{N} (1-\hat{s}_i) \cdot \hat{y}_i$, $n_1^- = \sum_{i=1}^{N} \hat{s}_i \cdot (1-\hat{y}_i)$, et $n_0^- = \sum_{i=1}^{N} (1-\hat{s}_i) \cdot (1-\hat{y}_i)$. Ces quatre nombres n_1^+, n_0^+, n_1^- et n_0^- sont les cardinalités des quatre groupes d'exemples définissant les quatre opérations possibles (*resp.,* (i), (ii), (iii) et (iv)) du point de vue de l'équité. Pour chaque groupe, nous trions et cumulons les scores de confiance associés à ces exemples pour obtenir les tableaux suivants : T_{1^+}, T_{0^+}, T_{1^-} et T_{0^-}. Par exemple, T_{1^+} contient les scores de confiance associés aux n_1^+ exemples prédits positivement par h et assignés au groupe 1 dans la reconstruction initiale de l'attaquant. $T_{1^+}[i]$ est la somme des i scores de confiance les plus bas au sein de ce groupe d'exemples. Ainsi, $T_{1^+}[i]$ est exactement le coût minimal pour changer la valeur de l'attribut sensible reconstruit de 1 à 0 pour i exemples prédits positivement par h. Nous utilisons quatre variables de décision entières, modélisant le nombre de fois où chacune des quatre opérations est effectuée pour corriger la reconstruction. Nous définissons alors notre modèle efficace de correction de reconstruction : $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$.

Données

- Cardinalités de la reconstruction initiale n_1^+, n_0^+, n_1^- et n_0^-.
- Tableaux des scores de confiance de l'attaquant, triés et cumulés T_{1^+}, T_{0^+}, T_{1^-} et T_{0^-}.
- Information de l'équité : h respecte une contrainte d'équité selon une métrique (*e.g.*, SP) et une valeur de tolérance ϵ

Variables de décision

- $s_{01}^+ \in [0, n_0^+]$: nombre de changements de $\hat{s}_i$ de 0 en 1, pour des exemples tels que $\hat{y}_i = 1$.
- $s_{10}^+ \in [0, n_1^+]$: nombre de changements de $\hat{s}_i$ de 1 en 0, pour des exemples tels que $\hat{y}_i = 1$.
- $s_{01}^- \in [0, n_0^-]$: nombre de changements de $\hat{s}_i$ de 0 en 1, pour des exemples tels que $\hat{y}_i = 0$.
- $s_{10}^- \in [0, n_1^-]$: nombre de changements de $\hat{s}_i$ de 1 en 0, pour des exemples tels que $\hat{y}_i = 0$.

Modèle $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$

$$\min T_{0^+}[s_{01}^+] + T_{1^+}[s_{10}^+] + T_{0^-}[s_{01}^-] + T_{1^-}[s_{10}^-] \quad (5)$$

$$s.t. : n_0^+ + n_0^- - s_{01}^+ - s_{01}^- + s_{10}^+ + s_{10}^- > 0 \quad (6)$$

$$n_1^+ + n_1^- - s_{10}^+ - s_{10}^- + s_{01}^+ + s_{01}^- > 0 \quad (7)$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{n_1^+ - s_{10}^+ + s_{01}^+}{n_1^+ + n_1^- - s_{10}^+ - s_{10}^- + s_{01}^+ + s_{01}^-} \leq \epsilon \quad (8)$$

$$-\epsilon \leq \frac{\sum_{i=1}^{N} \hat{y}_i}{N} - \frac{n_0^+ - s_{01}^+ + s_{10}^+}{n_0^+ + n_0^- - s_{01}^+ - s_{01}^- + s_{10}^+ + s_{10}^-} \leq \epsilon \quad (9)$$

Tout comme pour le modèle général, l'objectif (5) minimise les changements effectués (pondérés par la confiance de l'attaquant). Il peut être implémenté efficacement en utilisant des contraintes `element` au sein d'un solveur de Programmation par Contraintes (PPC). De telles contraintes sont utilisées pour accéder à un tableau de constantes à la position donnée par la valeur d'une variable : $T_{0^+}[s_{01}^+] = \texttt{element}(T_{0^+}, s_{01}^+)$. Par ailleurs, pour minimiser uniquement le nombre de changements, il est également possible de simplement sommer les quatre variables de décision. L'objectif devient alors linéaire et le modèle peut dans ce cas être résolu par un solveur PLNE générique. Les contraintes (6) et (7) s'assurent simplement que la reconstruction contienne au moins un exemple de chaque groupe protégé. Enfin, les contraintes (8) et (9) encodent la contrainte d'équité pour la métrique Statistical Parity. Plus généralement, $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ peut être utilisé pour encoder n'importe quelle contrainte de taux (utilisant les attributs sensibles) sur les prédictions du modèle cible h, ce qui inclue toutes les métriques d'équité statistique.

Une fois le modèle résolu, les affectations optimales des quatre variables de décision définissent les changements de coût minimal (pondérés par la confiance de l'attaquant) qui doivent être effectués pour rétablir l'équité. Dans une étape de post-traitement, les mouvements associés sont appliqués aux exemples correspondants, par ordre croissant des scores de confiance (de telle sorte que le coût global soit exactement la valeur de la fonction objectif (5) du modèle résolu). On obtient alors la version corrigée de la reconstruction S^*.

3.4 Généralisation des modèles

Les modèles décrits encodent directement la métrique d'équité Statistical Parity, mais peuvent également être utilisés pour corriger des reconstructions d'attributs sensibles à partir des autres métriques présentées dans la Table 1.

En effet, la Predictive Equality (PE) vise à égaliser les taux de Faux Positifs (entre les différents groupes protégés), ce qui est équivalent à la satisfaction de la Statistical Parity *sur le sous-ensemble des exemples négatifs* de l'ensemble d'entraînement. Ainsi, il est possible d'utiliser directement le modèle de correction de reconstruction proposé, en l'applicant uniquement sur ce sous-ensemble. En effet, la PE ne donne aucune information sur les exemples positifs. De la même manière, l'Equal Opportunity égalise les taux de Vrais Positifs, et la correction de reconstruction peut être faite en utilisant les modèles proposés *sur le sous-ensemble des exemples positifs* de l'ensemble d'entraînement. Enfin, pour la métrique Equalized Odds, il est possible d'appliquer successivement la méthode de correction de reconstruction de la Predictive Equality et celle de l'Equal Opportunity.

4 Evaluation expérimentale

Dans cette section, nous présentons une évaluation expérimentale de notre approche de correction de reconstruction. Nous considérons trois jeux de données de la littérature, présentant des caractéristiques (taille, attribut sensible, prédiction) variées, quatre métriques d'équité et de nombreuses valeurs de tolérance ϵ. Nous décrivons tout d'abord l'attaquant *baseline* considéré et la configuration de nos expériences avant de présenter les résultats obtenus.

4.1 Reconstruction *baseline*

Nous instancions le cadre décrit dans la Figure 1 avec un attaquant de la littérature, noté $\mathcal{A}'$ (implémentant le composant "attaquant *baseline*"). De manière consistante avec la littérature des attaques de reconstruction [11], nous considérons que le jeu de données contient une *"grande quantité d'information identifiante non-privée et un bit secret, un par individu."* [14]. Ici, le bit secret (privé) de chaque individu i est la valeur de son attribut sensible s_i. L'attaquant *baseline* $\mathcal{A}'$ connaît ainsi les attributs non sensibles de l'ensemble d'entraînement X et les labels Y (*i.e.*, toutes les colonnes de l'ensemble d'entraînement à l'exception de la colonne *secrète*, qui est celle de l'attribut sensible dans notre cas). Par ailleurs, $\mathcal{A}'$ a également accès à un *ensemble d'attaque*, $D_A = (X_A, S_A, Y_A)$ issu de la même distribution que l'ensemble d'entraînement (mais disjoint avec celui-ci). Cet ensemble d'attaque modélise la connaissance d'une approximation de la distribution des attributs sensibles par rapport aux attributs non sensibles et aux labels. En effet, l'utilisation d'un tel *ensemble d'attaque* pour entraîner un *modèle d'attaque* est cohérent avec la littérature [1].

L'attaquant $\mathcal{A}'$, proposé par [1], a donc accès à toute l'information dont notre correcteur de reconstruction aura besoin (en réalité, même davantage), ce qui constitue l'attaquant *baseline* le plus fort, que nous allons comparer avec notre correction de reconstruction. En résumé, $\mathcal{A}'$ a accès à un ensemble d'attaque $D_A = (X_A, S_A, Y_A)$, aux attributs non sensibles de l'ensemble d'entraînement X et à ses labels Y. Il a également un accès en boîte noire au modèle entraîné h, qui lui permet de connaître ses prédictions sur l'ensemble d'entraînement $\hat{Y} = h(X)$ et sur l'ensemble d'attaque $\hat{Y}_A = h(X_A)$. Cet ensemble d'attaque lui permet d'entraîner un modèle d'apprentissage à prédire S_A à partir de $(X_A, Y_A, \hat{Y}_A)$. Enfin, il utilise ce *modèle d'attaque* pour prédire $\hat{S}$ étant donné $(X, Y, \hat{Y})$.

4.2 Scores de confiance

Le *modèle d'attaque* de l'attaquant *baseline* effectuant une tâche de classification binaire (puisqu'on considère le cas des attributs sensibles binaires), ses scores de confiance se situent naturellement entre 0.5 et 1.0. Utiliser directement ces scores pour pondérer l'objectif de notre problème de correction de reconstruction signifierait donc que modifier une prédiction dont le score de confiance serait 1.0 (l'attaquant était certain de sa reconstruction) serait moins coûteux que de modifier deux prédictions avec un score de confiance de 0.51 (pour lesquelles l'attaquant n'était pas du tout sûr). Pour encourager la correction de reconstruction à se concentrer sur les prédictions associées aux scores de confiance les plus faibles, nous normalisons tous les scores de confiance avant de leur appliquer un même exposant $k \geq 1$ pour accroître leurs différences. En pratique, l'exposant k est choisi pour maximiser la qualité de la reconstruction obtenue sur un ensemble de validation $D'_A \subset D_A$. En résumé, l'attaquant $\mathcal{A}'$ produit une reconstruction $\hat{S} = \{\hat{s}_{i \in \{1 \dots N\}}\}$ des attributs sensibles de l'ensemble d'entraînement, accompagnée de scores de confiance $P = \{p_{i \in \{1 \dots N\}}\}$ normalisés et mis à la puissance k.

4.3 Configuration

Jeux de données Nous considérons trois jeux de données de tailles différentes, et sélectionnons un attribut sensible différent pour chacun d'eux afin d'obtenir des scénarios suffisamment variés. Le premier jeu de données utilisé est UCI Adult Income [12], un jeu de données très populaire dans la littérature de l'équité. Il rassemble des données sur le recensement de 1994 aux États-Unis, et la tâche associée est de prédire si une personne gagne plus de $50\,000$\$ par an. L'attribut sensible considéré est le genre (femme/homme). Nous utilisons également deux jeux de données construits à partir des résultats d'une enquête du bureau américain de recensement intitulée "American Community Survey (ACS) Public Use Microdata Sample (PUMS)". Plus précisément, ces jeux de données sont issus des données collectées dans l'état du Texas en 2018. Le second, nommé ACSPublicCoverage [10], contient des données sur des individus âgés de moins de 65 ans et ayant un revenu inférieur à $30\,000$\$, la tâche associée étant de prédire s'ils sont couverts par une assurance-maladie publique. L'âge sert ici d'attribut sensible (quartile le plus jeune/autres). Enfin, le troisième jeu de données, ACSIncome [10], rassemble des informations sur des individus âgés de plus de 16 ans, qui ont indiqué travailler au moins une heure par semaine l'année passée, pour un revenu d'au moins 100\$. Comme pour UCI Adult Income, la tâche de classification associée est de prédire si les individus gagnent plus ou moins de $50\,000$\$ par an. Nous utilisons une version binarisée de l'origine ethnique ("blancs"/autres) comme attribut sensible.

Ces informations sont synthétisées dans la Table 2. Pour toutes les expériences, chaque jeu de données est partagé entre un ensemble d'entraînement $\left(\frac{1}{3}\right)$, un ensemble de test $\left(\frac{1}{3}\right)$ et un ensemble d'attaque $\left(\frac{1}{3}\right)$. Ici, l'ensemble de test sert uniquement à s'assurer que le modèle équitable a été correctement entraîné (en particulier, à montrer qu'il n'a pas *sur-appris*). L'ensemble d'attaque est connu par l'attaquant *baseline* (comme décrit dans la section 4.1).

Modèles équitables (cibles) Nous utilisons une méthode populaire d'apprentissage équitable, implémentée dans la librairie `Fairlearn` [6]. Cette méthode in-processing, nommée ExponentiatedGradient [2], formule le problème de classification équitable comme une séquence de problèmes de classification pondérée. Etant donné un modèle de base sensible aux coûts, l'approche consiste en un jeu à deux joueurs dans lequel un joueur entraîne le modèle de base tandis que l'autre adapte les poids des exemples d'entraînement. Nous utilisons des arbres de décision de la librarie `scikit-learn` [24] comme modèles de base, avec une profondeur maximale fixée à 8 et tous les autres paramètres laissés à leur valeur par défaut. Notons cependant que notre approche est agnostique au type d'algorithme d'apprentissage équitable mis en oeuvre, puisqu'elle utilise seulement l'information sur l'équité du modèle final.

Métriques d'équité Nous effectuons nos expériences pour les quatre métriques présentées dans la Table 1. Nous utilisons 49 valeurs différentes de tolérance d'inéquité ϵ, variant de manière non linéaire entre 0.0 (équité parfaite) et 0.20.

TABLE 2 – Synthèse des jeux de données utilisés dans nos expériences

Ref.	Nom	Tâche de prédiction (binaire)	#Exemples	#Attributs non sensibles	Attribut sensible
[12]	UCI Adult Income	Revenu > 50K$	45 222	7 catégoriques, 6 numériques	Genre (Femme/Homme)
[10]	ACSPublicCoverage*	Couvert par une assurance maladie publique	98 928	17 catégoriques, 1 numérique	Age (Premier Quartile/Autres)
[10]	ACSIncome*	Revenu > 50K$	135 924	7 catégoriques, 2 numériques	"Code" ethnique (Blancs/Autres)

*(État du Texas, 2018)

Modèles d'attaque Les modèles d'attaque utilisés par l'attaquant *baseline* $\mathcal{A}'$ sont des forêts aléatoires de la librarie `scikit-learn` [24]. Les attributs sensibles étant souvent déséquilibrés [1], nous utilisons une fonction objectif classe-pondérée. Les hyperparamètres des forêts aléatoires sont optimisés par la librarie `HyperOpt-Sklearn` [22], avec un maximum de 100 itérations pour son algorithme de recherche "Tree of Parzen Estimators". Cette configuration permet de s'assurer que l'attaquant *baseline* représente une base solide face à notre étape de correction de reconstruction, et correspond aux pratiques de la littérature.

Correction de reconstruction Notre modèle efficace de correction de reconstruction $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$ (décrit dans la Section 3.3) est implémenté et résolu par le solveur commercial `IBM ILOG CP Optimizer Version 12.10` via l'API Python de modélisation `DOcplex` (version 2.21.207) dans sa configuration par défaut. Le nombre de threads utilisés par CP Optimizer est fixé à 1 et la tolérance d'optimalité (absolue et relative) est mise à 0.0. En effet, en raison du processus d'exponentiation présenté en Section 4.1, les valeurs des scores de confiance peuvent être très petites et seraient inférieures à la tolérance d'optimalité par défaut du solveur. Notre méthode de correction de reconstruction est implémentée comme une classe Python et est disponible sur notre dépôt [1].

Paramètres expérimentaux Nous fixons un temps d'exécution maximum d'une minute pour l'étape de correction de reconstruction (création et résolution du modèle). En pratique, cette limite n'a jamais été atteinte, et tous les modèles ont été résolus à l'optimum en quelques secondes (moins d'une seconde en moyenne). Chaque expérience est répétée 100 fois, avec différents paramètres pour l'initialisation des générateurs pseudo-aléatoires (pour la séparation du jeu de données et l'initialisation des algorithmes). Les résultats sont moyennés sur les 100 exécutions, et les déviations standard sont reportées. Toutes les expériences sont exécutées sur une plateforme de calcul équipée de processeurs Intel Xeon E5-2683 v4 Broadwell @ 2.1GHz CPU.

4.4 Résultats

Les résultats de nos expériences sont présentés pour les trois jeux de données et les quatre métriques d'équité considérés dans les Figures 2a, 2b et 2c. Ils démontrent l'efficacité de l'approche proposée. Comme mentionné en Section 4.1, l'attaquant $\mathcal{A}'$ utilise déjà toute l'information dont dispose notre composant de correction de reconstruction. Ainsi, toute amélioration de la reconstruction par notre processus de correction peut uniquement être expliquée par la sémantique de la contrainte d'équité intégrée explicitement dans notre modèle $\mathcal{RC}_{\mathcal{E}}(\hat{S}, P, \hat{Y}, \epsilon)$. Pour rappel, la précision de la reconstruction est la proportion d'exemples d'entraînement e_i pour lesquels l'attribut sensible $s_i \in S$ a été correctement reconstruit (par l'attaquant *baseline* dans $\hat{s}_i \in \hat{S}$ ou dans la reconstruction corrigée $s_i^* \in S^*$).

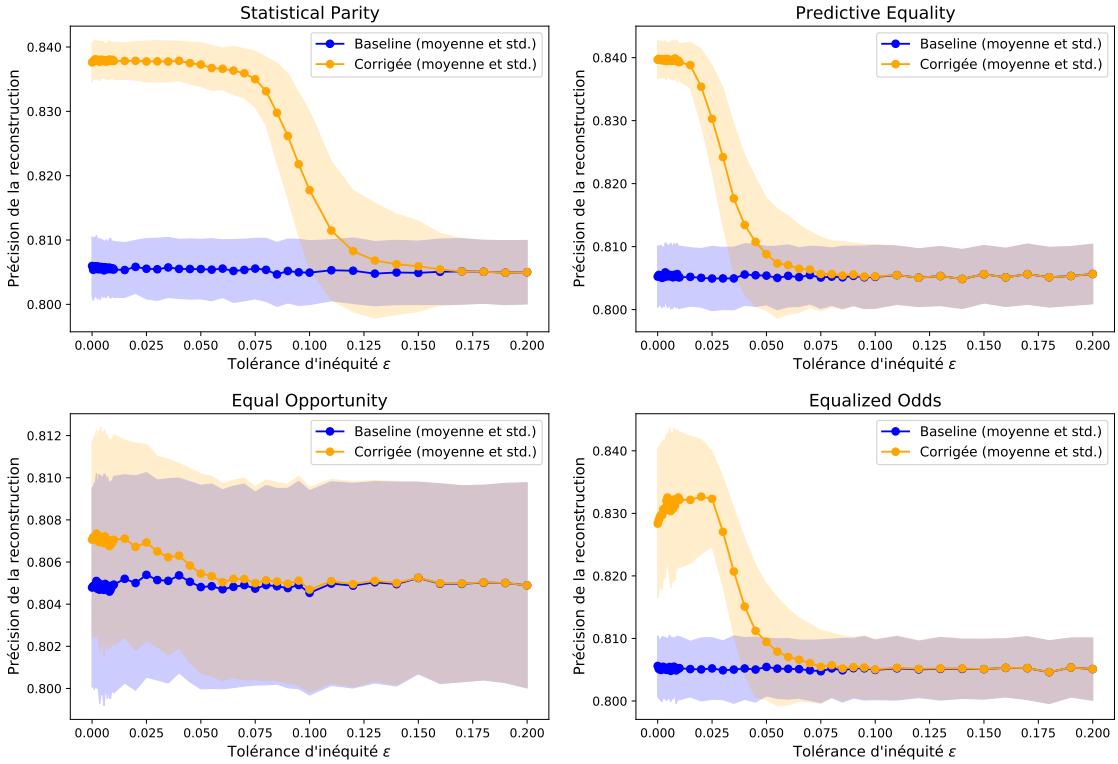
On peut observer que la reconstruction corrigée est toujours plus précise que la reconstruction originale de l'attaquant *baseline*, ce qui indique que les changements effectués sont corrects la plupart du temps. Par ailleurs, plus la contrainte d'équité utilisée est forte (*i.e.*, petites valeurs de tolérance d'inéquité ϵ), plus l'amélioration permise par la correction de reconstruction est importante. En effet, cette dernière est liée à la quantité de biais atténuée par la méthode d'apprentissage équitable, qui à son tour dépend de la métrique d'équité considérée, de la tolérance d'inéquité et du biais des données d'origine. Pour des contraintes d'équité fortes, on observe des améliorations dans la précision de la reconstruction allant jusqu'à 0.06 (ou 9%), comme dans le cas de la métrique Statistical Parity avec le jeu de données ACSIncome. De telles améliorations sont uniquement dues à l'information de l'équité, qui est la seule contrainte dans notre modèle de correction de reconstruction.

Pour rappel, la métrique Predictive Equality (*resp.,* Equal Opportunity) s'applique uniquement aux exemples négatifs (*resp.,* positifs). Ces métriques ne peuvent donc corriger qu'une partie de la reconstruction *baseline* (*cf.* Section 3.4). Les jeux de données utilisés étant déséquilibrés, avec une majorité d'exemples négatifs, la métrique Equal Opportunity ne s'applique donc qu'à une minorité d'exemples. Par conséquent, les améliorations de reconstruction observées avec cette métrique sont plus modestes que pour les autres. En effet, même avec un taux de modifications correctes proche, le nombre de corrections effectuées est plus faible.

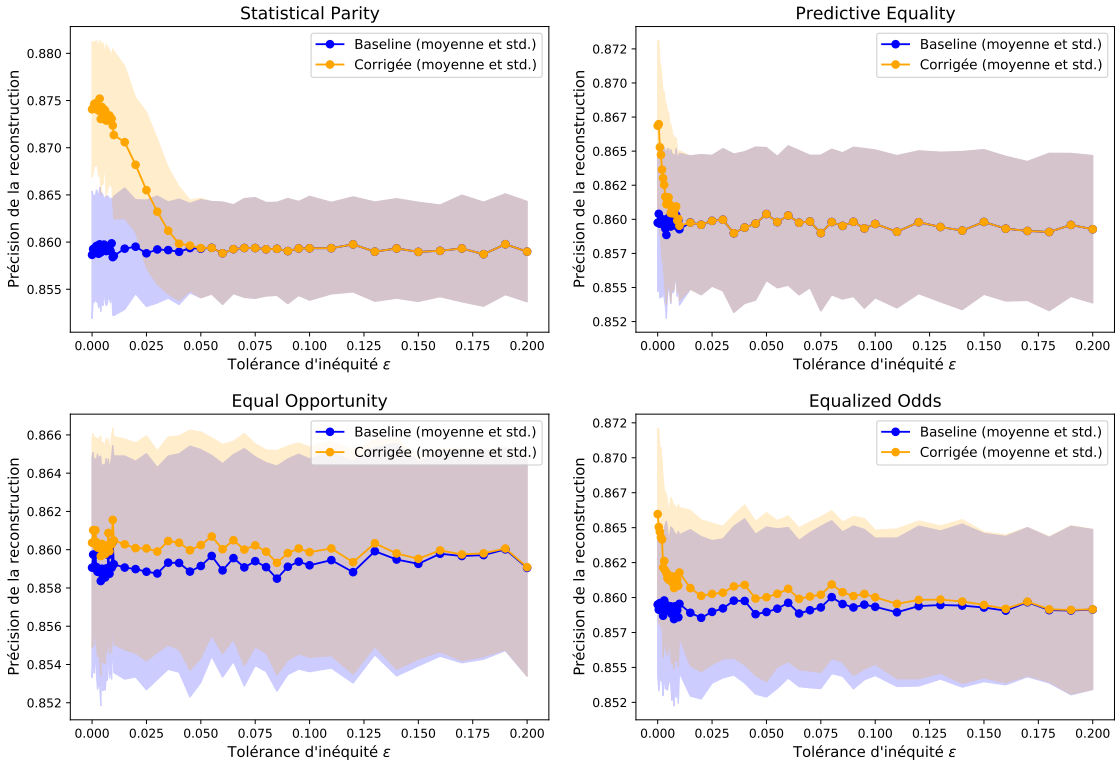
En variant la tolérance d'inéquité ϵ, la seule entrée de la méthode de reconstruction *baseline* qui est modifiée est le vecteur des prédictions du modèle équitable $\hat{Y}$ (et l'information de l'équité). Le fait que la précision de la reconstruction *baseline* de $\mathcal{A}'$ soit quasiment constante lorsque ϵ varie montre que les prédictions du modèle équitable sont peu utilisées par les modèles d'attaque construits. A l'inverse, notre méthode de correction sait exactement comment interpréter l'information de l'équité vis-à-vis de $\hat{Y}$, et est capable de l'utiliser pleinement pour améliorer la qualité de la reconstruction.

Finalement, les résultats expérimentaux montrent que notre méthode de correction de reconstruction est capable d'améliorer significativement la reconstruction des attributs sensibles de l'ensemble d'entraînement d'un modèle, même par rapport à un attaquant *baseline* aussi informé.

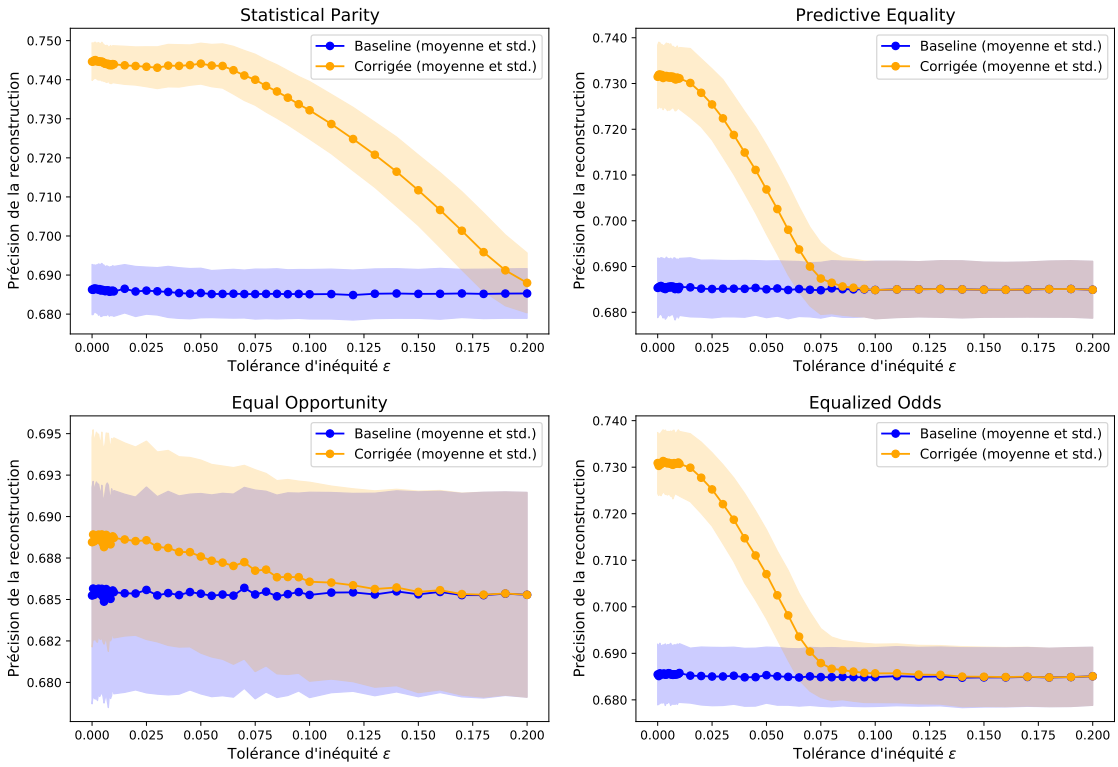
1. `https://github.com/ferryjul/SensitiveAttributesReconstructionCorrector/`



(a) Jeu de données UCI Adult Income.



(b) Jeu de données ACSPublicCoverage.



(c) Jeu de données ACSIncome.

FIGURE 2 – Qualité des reconstructions *baselines* (attaquant $\mathcal{A}'$) et corrigées, pour nos différentes expériences.

5 Discussion

Dans ce travail, nous proposons une approche novatrice utilisant la programmation déclarative pour améliorer les performances de reconstruction de n'importe quel attaquant *baseline*, en incorporant des contraintes définies par l'utilisateur. Bien que le problème général soit difficile d'un point de vue calculatoire, nous montrons que dans le cas des métriques d'équité statistique (ou d'autres contraintes formulées au niveau des groupes), il peut être reformulé et résolu de manière très efficace. Par ailleurs, notre large étude expérimentale montre que, parce qu'ils utilisent les attributs sensibles pour s'assurer de l'équité des modèles construits, les algorithmes d'apprentissage équitable divulguent intrinsèquement des informations sur ceux-ci. En effet, les contraintes d'équité fournissent des informations sur la distribution des prédictions (sur l'ensemble d'entraînement) d'un modèle équitable par rapport aux attributs sensibles (de l'ensemble d'entraînement). Même si cette information se situe au niveau des groupes, elle peut être utilisée par un attaquant pour améliorer sa reconstruction *baseline* des attributs sensibles. Par ailleurs, plus la contrainte d'équité appliquée est forte, plus les améliorations des reconstructions observées en pratique sont importantes.

Les travaux futurs incluent la combinaison de notre approche avec différents attaquants *baselines*, l'optimisation du traitement des scores de confiance P, ainsi que l'utilisation de notre méthode dans le contexte plus général des attributs sensibles non binaires. Enfin, tirer profit de la nature déclarative de l'étape de correction de reconstruction pour intégrer d'autres types de contraintes est également une direction de recherche intéressante.

Notre travail complet [17] contient un certain nombre de contributions que nous n'avons pas pu présenter ici pour des raisons d'espace. Nous présentons notamment un état de l'art plus détaillé, ainsi que des résultats expérimentaux supplémentaires, incluant (i) les performances des modèles équitables (cibles) appris en termes d'équité et de précision, (ii) les résultats de nos expériences utilisant un autre attaquant *baseline*, moins informé que $\mathcal{A}'$ et (iii) les résultats de nos expériences attaquant d'autres modèles équitables, entraînés avec des approches de pre-processing ou de post-processing. Nous discutons des contre-mesures possibles et montrons notamment que même si l'information sur l'équité du modèle n'est pas révélée, un attaquant peut adopter des stratégies relativement simples pour l'estimer. La correction de reconstruction avec la contrainte estimée présente alors de bonnes performances. Cela démontre l'applicabilité de l'approche proposée, et suggère que les métriques d'équité statistique, puisqu'elles utilisent explicitement les attributs sensibles, entrent intrinsèquement en conflit avec la protection de ces derniers. Nous démontrons par ailleurs que les deux modèles proposés (modèle général et modèle efficace) partagent le même ensemble de solutions optimales lorsque les contraintes considérées sont des contraintes d'équité statistique. Enfin, nous discutons de l'extension de ces modèles au cas général des attributs sensibles multi-valués.

Références

[1] Jan Aalmoes, Vasisht Duddu, and Antoine Boutet. Dikaios : Privacy auditing of algorithmic fairness via attribute inference attacks. *arXiv preprint arXiv :2202.02242*, 2022.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 2018.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[4] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3) :671–732, 2016.

[5] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, et al. AI fairness 360 : An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018.

[6] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, et al. Fairlearn : A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[7] Simon Caton and Christian Haas. Fairness in machine learning : A survey. *arXiv preprint arXiv :2010.04053*, 2020.

[8] Alexandra Chouldechova. Fair prediction with disparate impact : A study of bias in recidivism prediction instruments. *Big data*, 5(2) :153–163, 2017.

[9] Emiliano De Cristofaro. An overview of privacy in machine learning. *CoRR*, abs/2005.08679, 2020.

[10] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult : New datasets for fair machine learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34 : Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 6478–6490, 2021.

[11] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In Frank Neven, Catriel Beeri, and Tova Milo, editors, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210. ACM, 2003.

[12] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[14] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed ! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4(1) :61–84, 2017.

[15] The U.S. EEOC. Uniform guidelines on employee selection procedures. March 2, 1979.

[16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, et al., editors, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268. ACM, 2015.

[17] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Exploiting fairness to enhance sensitive attributes reconstruction. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

[18] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks : A survey. *CoRR*, abs/2202.08187, 2022.

[19] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[20] Hui Hu and Chao Lan. Inference attack and defense on the distributed private fair learning framework. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2020.

[21] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1) :1–33, 2012.

[22] Brent Komer, James Bergstra, and Chris Eliasmith. Hyperopt-sklearn : automatic hyperparameter configuration for scikit-learn. In *ICML workshop on AutoML*, volume 9, page 50. Citeseer, 2014.

[23] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6) :115 :1–115 :35, 2021.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, et al. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.