MAP 511 : RAPPORT D'EA DE RECHERCHE

Étude de l'algorithme COMPAS dans la controverse opposant ProPublica à Northpointe

Sous la direction de Jean-Michel LOUBES

12 décembre 2023

Antoine MARTINEZ, Céleste GALLIEN, Swann BESSA

ÉCOLE POLYTECHNIQUE

IP PARIS

TABLE DES MATIÈRES

INTRODUCTION

L'avènement de l'ère numérique a engendré une révolution dans de nombreux domaines avec le développement de l'intelligence artificielle et des algorithmes prédictifs. Salués tout d'abord comme des outils révolutionnaires, permettant d'apporter objectivité et impartialité, ils soulèvent néanmoins une question importante : dans quelle mesure sont-ils immunisés contre les biais et les discriminations qui existent dans notre société ? Peuvent-ils amplifier les inégalités déjà existantes ?

Cette problématique trouve une résonance particulière dans le domaine de la justice, où des décisions éclairées et impartiales sont cruciales, et soulève des enjeux complexes, mis en lumière par des cas comme la controverse Northpointe – ProPublica, datant de 2016, sur laquelle nous avons travaillé au cours de notre EA.

Au cœur de cette controverse se trouve l'algorithme COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), outil utilisé dans le système judicaire américain pour évaluer le risque de récidive des individus condamnés (sous forme d'un score allant de 1 à 10, 10 étant « à haut risque de récidive »). Le contentieux a opposé Northpointe, entreprise à l'origine de COMPAS, à ProPublica, un média d'investigation, qui a exposé des tendances discriminatoires de l'algorithme, biaisé contre les Afro-Américains. Ce cas a permis de mettre en avant les limites de la prétendue neutralité des algorithmes, dans le cas spécifique du domaine judicaire, domaine particulièrement sensible.

Ce cas soulève donc de nombreuses questions en termes de machine learning allant de la qualité des données d'entraînement, qui peuvent refléter des disparités sociales existantes, aux méthodes de correction des biais dans les algorithmes, en passant par la réflexion sur quels critères de « fairness » à prendre en compte pour quantifier ces biais.

Le rapport suivant s'articulera en cinq parties principales. Après avoir rappelé le contexte de l'affaire et les résultats déjà existants, nous ferons un état de l'art de ce qui existe en matière de fairness, des critères à prendre en compte et des méthodes de correction des biais, appliquées à l'algorithme COMPAS. Puis, nous analyserons plus spécifiquement cet algorithme afin d'ajouter des éléments d'analyse supplémentaires par rapport à l'étude originale. Nous présenterons ensuite notre démarche de réalisation d'un nouvel algorithme, plus performant et diminuant les biais par rapport à l'algorithme original. Enfin, nous terminerons ce rapport par une réflexion éthique et juridique sur ce type de biais, analyse motivée par la réalisation d'une plaidoirie devant une magistrate, au cours de notre EA.

1 CONTEXTE

1.1 Définition des termes

Le dataset sur lequel nous avons travaillé, qui est le même que celui utilisé par ProPublica, contient les données de plus de 7000 personnes (environ 5200 lorsqu'on restreint aux groupes des Afro-Américains et des Caucasiens) du comté de Broward en Floride, qui avaient été évaluées par l'algorithme COMPAS. Ces données comprennent des informations sur les antécédents criminels des individus, leur âge, leur origine ethnique, les scores de risque attribués par l'algorithme et si ces individus ont récidivé ou non deux ans après leur libération. Rappelons ici que le score est un nombre entre 1 et 10 (de 1 à 4 étant « bas risque » et de 5 à 10 étant « haut risque »).

La controverse entre Northpointe et ProPublica, que nous allons exposer par la suite, repose en grande partie sur la mesure des biais de l'algorithme COMPAS à l'aide de deux critères distincts. Présentons ici ces critères.

ProPublica : Afin de mesurer le biais dans l'algorithme, ProPublica compare, pour chaque groupe (les Afro-Américains et les Caucasiens), la « probabilité du score sachant la récidive » :

$\mathbb{P}$ (être catégorisé haut risque | a récidivé) et $\mathbb{P}$ (être catégorisé haut risque | n'a pas récidivé)

Le « critère ProPublica » correspond à celui dit « *equality of odds* ». Dans la définition ci-dessous, Y est la réalité c'est-à-dire le statut de récidive ($Y = 1$ correspond à l'individu a récidivé, $Y = 0$ est l'individu n'a pas récidivé). $\hat{Y}$ est la prédiction, ici le score ($\hat{Y} = 1$ correspond à l'individu a été classifié haut risque, $\hat{Y} = 0$ est l'individu a été classifié bas risque). Enfin, S correspond à la variable sensible, ici l'origine ethnique ($S = 0$ correspond aux Afro-Américains, $S = 1$ aux Caucasiens).

$$EoO(\hat{Y}, S) = \frac{\mathbb{P}(\hat{Y} = 1 | Y = 1, S = 0)}{\mathbb{P}(\hat{Y} = 1 | Y = 1, S = 1)}$$

Plus cette valeur est proche de 1 et plus un algorithme est considéré comme non biaisé, par rapport à la variable S.

Northpointe : Northpointe utilise un critère « inversé » par rapport à celui de ProPublica et compare, pour chaque groupe, la « probabilité de la récidive sachant le score », i.e. :

$\mathbb{P}$ (ne pas avoir récidivé | catégorisé haut risque) et $\mathbb{P}$ (avoir récidivé | catégorisé bas risque)

Ce critère correspond à celui dit de « *predictive parity* », c'est-à-dire l'égalité de la FPP et la FNP entre les deux groupes où la FPP est la false positive prediction (parmi ceux classés haut

risque, ceux qui n'ont pas récidivé) et la FNP est la false negative prediction (parmi ceux classés bas risque, ceux qui ont récidivé). Une fois ces deux critères présentés, exposons la controverse entre Northpointe et ProPublica.

1.2 Les résultats de Propublica

En mai 2016, ProPublica publie « *Machine Bias* » par Julia Angwin, Jeff Larson, Surya Mattu et Lauren Kirchner [1], article qui met en évidence des disparités de traitement entre les blancs et les Afro-Américains par l'algorithme COMPAS, ce qui suggère de fort préjugés raciaux. Nous avons reproduit l'analyse de ProPublica et les chiffres et graphiques présentés ci-dessous sont les résultats que nous avons obtenus (très similaires à ceux de ProPublica).

Tout d'abord, en représentant la distribution des scores pour chaque groupe, on observe que celle-ci est très différente avec davantage de scores élevés pour les Afro-Américains.

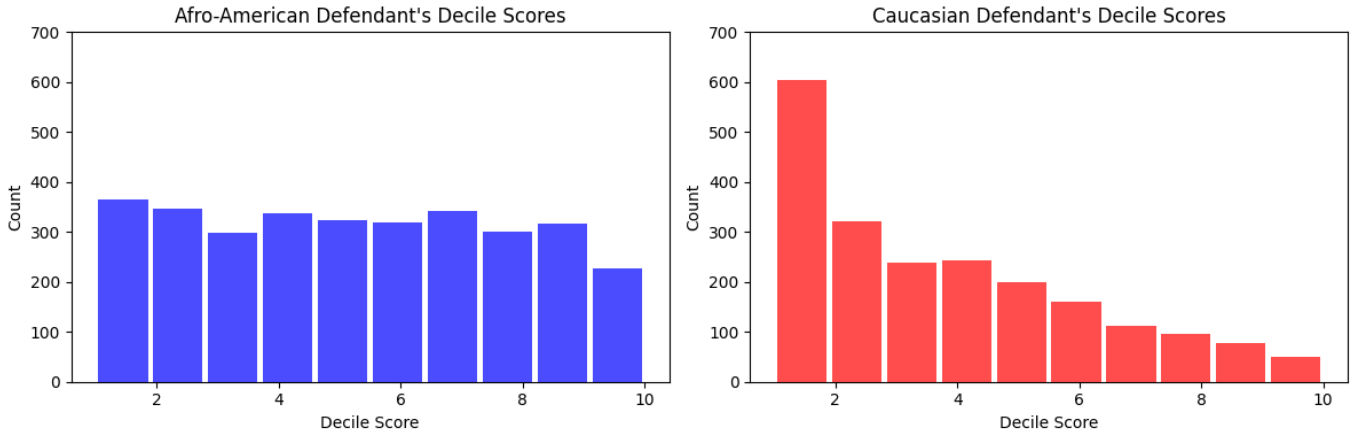

Figure 1 – Distribution des scores pour les blancs et les Afro-Américains

Par ailleurs, cette analyse révèle le résultat principal suivant : lorsqu'il s'agit de prévoir qui va récidiver, l'algorithme commet des erreurs avec les prévenus noirs et blancs, à peu près au même rythme, mais de manière très différente.

- L'algorithme est particulièrement susceptible d'identifier à tort les accusés noirs comme de futurs criminels. En effet, parmi toutes les personnes qui n'ont pas récidivé, les prévenus noirs sont presque deux fois plus susceptibles que d'être classés « haut risque » que les prévenus blancs (23.5% vs. 44.9%)
- À l'inverse, les prévenus blancs sont étiquetés à tort comme présentant un faible risque plus souvent que les prévenus noirs. En effet, parmi toutes les personnes qui ont récidivé, les prévenus blancs sont plus susceptibles d'être classés « bas risque » que les prévenus noirs (47.7% vs. 28%).

Pour vérifier ces résultats, nous avons recalculé le taux de faux positifs (haut risque, pas de récidive) et de faux négatifs (bas risque, récidive) pour chaque groupe et avons trouvé des

résultats très similaires, aux erreurs près (cf. tableau ci-dessous).

Résultats	Blancs	Afro-Américains
% de faux positifs	22.0%	42.3%
% de faux négatifs	49.6%	28.5%

L'analyse effectuée par ProPublica repose donc sur le critère d'*equality of odds* présenté ci-dessus. Lorsqu'on le calcule, celui-ci vaut 1,42 et est donc très loin de 1 et donc d'une situation d'égalité. En effet, la probabilité d'avoir un score élevé sachant qu'on a récidivé est 1,42 fois plus élevée pour un Afro-Américain que pour un blanc. Regardons maintenant la réponse de Northpointe à cette analyse.

1.3 La réponse de Northpointe

À cet article, la compagnie Northpointe a publié une réponse en juillet 2016 « *COMPAS Risk Scales : Demonstrating accuracy equity and predictive parity* » [2] dans laquelle ils disent rejeter fermement les accusations de biais raciaux de ProPublica. Leur argument principal est que la mesure de fairness utilisée par ProPublica n'est pas adaptée à ce cas et est irréaliste car elle ne prend pas en compte la différence du taux de récidive entre blancs et Afro-Américains. Selon eux, la métrique à utiliser dans ce cas est celle de *predictive parity* : COMPAS est juste si un score a la même signification quelle que soit l'origine ethnique (la probabilité de récidiver sachant le score est la même pour chaque groupe). En utilisant ce critère, ils montrent qu'il n'existe pas différence de traitement, ce que nous avons résumé dans le tableau ci-dessous

Résultats	Blancs	Afro-Américains
% d'individus ayant récidivé parmi ceux classés à faible risque	29.0%	35.2%
% d'individus n'ayant pas récidivé parmi ceux classés à haut risque	40.5%	35.1%

Ainsi, avec cette mesure (et aux erreurs près), l'algorithme semble beaucoup plus juste et sans biais sur l'origine ethnique. Plusieurs questions se posent alors : quel critère utiliser pour mesurer la fairness d'un algorithme ? Comment réduire les biais ? Peut-on faire mieux que COMPAS et apporter des éléments supplémentaires d'analyse ? C'est à ces différentes questions que nous avons cherché à répondre pendant notre EA et que nous détaillons dans la suite du rapport, une fois le contexte de la controverse établi.

2 TRAVAUX RELIÉS

2.1 QUELS CRITÈRES EN MATIÈRE DE FAIRNESS ?

Avant de parler de *fairness*, il est nécessaire de définir le cadre dans lequel on se place. La littérature distingue trois types d'indices de *fairness* [3] :

- **Des critères de groupe** : Ce sont des statistiques calculées à l'échelle de chaque groupe (chaque groupe étant défini par une valeur de la variable sensible S), puis comparées entre elles. L'algorithme sera alors défini comme *fair* si ces critères ont les mêmes valeurs entre les groupes. Une revue des critères de groupes principalement utilisés dans la littérature [4] permet de montrer qu'ils sont en grande majorité basés sur les nombres de FP (*false positive*), TP (*true positive*), FN (*false negative*) et TN (*true negative*).
- **Des critères individuels** : Ces critères étudient la *fairness* à l'échelle individuelle. Basés sur des définitions topologiques [5], ils s'attachent à déterminer si deux individus proches en inputs obtiennent des outputs similaires.
- **Des définitions basées sur la causalité** : Ces définitions s'attachent à déterminer l'influence d'une modification artificielle de la variable sensible sur l'output.

Le désaccord entre ProPublica et NorthPointe porte sur l'opposition entre deux types de critères de groupes résumés dans le tableau suivant (dans le cas d'une classification binaire) :

Mesure	Formule	Interprétation Probabiliste
True Positive Rate, TPR	$\frac{TP}{TP+FN}$	$P(\hat{Y}=1\|Y=1,S)$
True Negative Rate, TNR	$\frac{TN}{TN+FP}$	$P(\hat{Y}=1\|Y=0,S)$
Positive Predictive Value, PPV	$\frac{TP}{TP+FP}$	$P(Y=1\|\hat{Y}=1,S)$
Negative Predictive Value, NPV	$\frac{TN}{TN+FN}$	$P(Y=1\|\hat{Y}=0,S)$

Ces critères de groupes sont les principaux utilisés dans la littérature, en plus d'un $5^{ème}$ critère non mentionné dans l'argumentation des deux parties, certainement car moins précis :

Disparate Impact, DI	$\frac{TP+FP}{TP+FP+TN+FN}$	$P(\hat{Y}=1\|S)$

Pour rappel, ProPublica a démontré que les deux premiers critères (TPR et TNR) étaient largement différents entre les Afro-Américains et les Blancs dans l'algorithme de Northpointe (c'est l'*equality of odds*). Dans sa réponse, Northpointe oppose les deux derniers critères (PPV et NPV), qui sont similaires pour les Afro-Américains et les Blancs (c'est la *predictive parity*).

Un argument en faveur de ProPublica peut être basé sur une notion juridique d'égalité. En effet, le principle d'égalité stipule que deux individus avec des caractéristiques similaires doivent être traités de la même façon par l'algorithme. Ainsi, comme les TPR et TNR conditionnent par rapport à la variable **réelle** Y (dans la définition probabiliste), ces critères correspondent a priori à une mesure d'égalité juridique. En revanche, dans le cas des PPV et NPV, le conditionnement est fait par rapport à la variable de score $\hat{Y}$. Comme $\hat{Y}$ est une variable **décidée par l'algorithme**, il est difficile de lier les critères PPV et NPV à un concept juridique de fairness. D'après nos connaissances, un tel lien n'a pas été réalisé dans la littérature et interroge donc sur la pertinence de ce critère en fairness.

Il est en tout cas indéniable que ces deux critères sont largement utilisés dans la littérature, et nous partirons donc du principe qu'ils sont tout les deux valables d'un point de vue juridique.

2.2 Incompatibilité des critères

Notre problématique consiste à déterminer si l'algorithme COMPAS est satisfaisant (voire optimal) en termes de *fairness*, et donc s'il a sa place dans le système juridique américain. Comme COMPAS vérifie deux des quatre mesures ci-dessus, la question se pose alors de savoir s'il est mathématiquement possible de vérifier simultanément les quatre critères, ou du moins de les optimiser. Auquel cas, COMPAS serait sous-optimal.

De l'étude de la littérature émerge un « théorème d'incompatibilité » entre divers critères de fairness montrant qu'ils ne sont pas tous vérifiables en même temps. En effet, Garg et al. [6] montrent que la satisfaction des critères TPR, TNR et PPV implique que $P(Y = 1|S = 0) = P(Y = 1|S = 1)$ (égalité des taux de récidive). Cela est démontré par un calcul probabiliste élémentaire exprimant $P(Y = 1)$ en fonction de ces critères :

$$P(Y = 1|S = 0) = \frac{PPV_0 FPR_0}{PPV_0 FPR_0 + (1 - PPV_0)TPR_0}$$

$$P(Y = 1|S = 1) = \frac{PPV_1 FPR_1}{PPV_1 FPR_1 + (1 - PPV_1)TPR_1}$$

Autrement dit, si les critères TPR, FPR et PPV sont les mêmes, comme le taux de récidive s'exprime en fonction de ces grandeurs, alors les taux de récidive seront les mêmes entre Blancs et Afro-Américains.

Or, dans notre cas, les taux de récidive sont différents (52% pour les Afro-Américains et 39% pour les blancs) et rendent donc impossible une conciliation parfaite des quatre critères cités en 2.1. Notre étude s'attache donc à optimiser ces critères de fairness en ayant conscience qu'un respect exact n'est pas possible.

Nous tenons à souligner que les articles sur la satisfaction de critères de fairness omettent l'accuracy de l'algorithme comme critère dans leurs études. Autrement dit, des situations a

priori « fair » avec un taux de précision de 50% (exemple d'un classifieur aléatoire) ne sont pas distinguées d'approches réalistes tenant compte d'un tradeoff entre *accuracy* et fairness.

2.3 Démarches antérieures pour corriger le biais

Nous étudions désormais les méthodes utilisées jusqu'à présent dans l'entraînement de modèles de machine learning pour garantir la fairness.

Une première idée intuitive serait de supprimer la variable sensible du dataset, afin que l'algorithme ne l'utilise pas pour prendre sa décision. Or, cette approche ne fonctionne pas car l'algorithme « réapprend » la variable sensible à partir des autres variables qui lui sont corrélées. De plus, cela pourrait poser un problème car le fait de la supprimer du dataset ferait qu'il ne serait plus possible de corriger le biais. Il faut donc trouver d'autres manières de corriger ce biais.

Tout d'abord, Besse & al. (dont notre encadrant Jean-Michel Loubes) [7] ont montré le succès de deux démarches distinctes (prédiction d'un revenu au-dessus ou en-dessous d'un certain seuil). Les deux méthodes sont :

- L'entraînement de classifieurs différents par variable sensible. Cette démarche consiste à diviser le dataset en plusieurs datasets, chacun correspondant à une valeur de la variable sensible. Alors, des classifieurs sont entrainés séparément sur chacun de ces datasets.
- L'utilisation de paliers différents (dans la classification binaire). Un seul classifieur est utilisé en entraînement, mais la probabilité en output est discrétisé différemment selon les groupes.

On notera que la première démarche pose problème dans un cas où le nombre de points est limité, car utiliser des classifieurs différents divise le dataset pour chaque classifieur.

Une autre méthode consiste à ajouter les critères de fairness comme des pénalités afin de les optimiser. Par exemple, nous pouvons citer Scutari & al. [8] qui utilisent une penalité ridge sur les variables sensibles ou encore Padala & Gujar [9] qui utilisent des réseaux de neurones pour optimiser des pénalités non-convexes correspondant aux critères cités en 2.1.

Enfin, ces deux dernières années, des méthodes basées sur un reweighting de la fonction de perte [10] ont montré leur utilité en particulier dans un cas où le dataset initial présente des disparités dans les distributions de paramètres. Alors que les approches précédentes conservaient une loss function souffrant de l'hétérogéneité du dataset, quitte à rajouter une autre pénalité, le reweighting a pour avantage de corriger d'emblée l'hétérogénéité du dataset initial.

3
ANALYSE DU DATASET ET DU BIAIS DANS L'ALGORITHME

3.1 Analyse de l'influence des inputs

Une fois l'étude de l'état de l'art en matière de fairness effectuée, nous nous sommes intéressés plus précisément à l'algorithme COMPAS. Nous avons essayé en premier lieu d'apporter des éléments nouveaux par rapport aux analyses effectuées par ProPublica et Northpointe (cf. partie 1), notamment dans la perspective de « rejouer » le procès devant une magistrate.

Au premier abord, lorsqu'on regarde la répartition des scores en fonction de l'ethnicité, comme le montre la figure 1, le modèle semble être complètement biaisé. Son utilisation dans certains états des États-Unis semble même absurde. Nous avons donc essayé de regarder l'influence des autres variables pour savoir si certaines pourraient expliquer en partie la distribution des scores. La première étape fut de réaliser un tableau de corrélation pour se donner une première idée des inputs qui pourraient nous intéresser. Nous avons pu constater que l'âge est assez fortement corrélé avec le score (-0.4), et nous avons donc décidé de nous y intéresser de plus près.

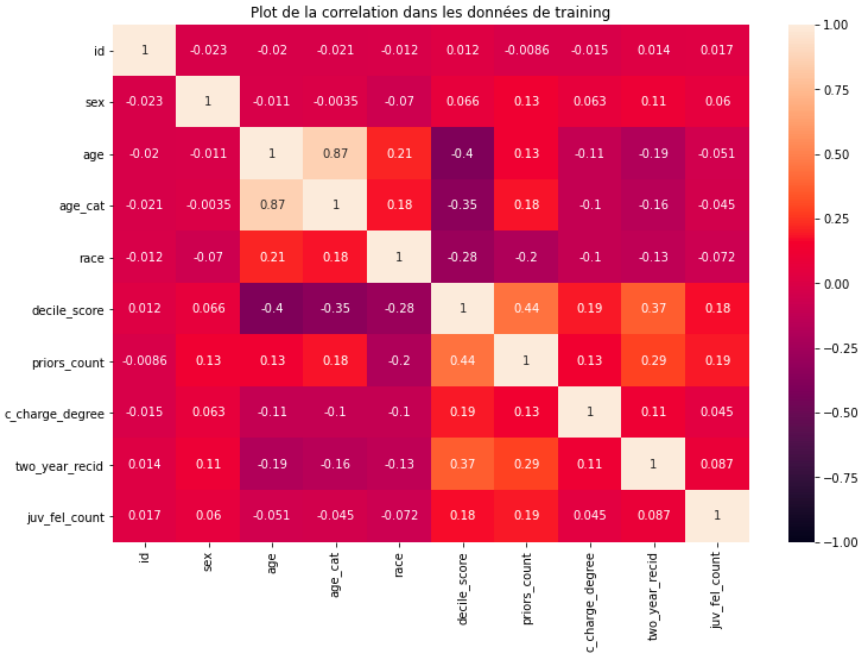

Figure 2 – Tableau de corrélation

Même si nous avons accès aux résultats de l'algorithme COMPAS, avec les variables decile score et score, nous n'avons pas accès directement à l'algorithme, donc il est difficile de savoir exactement quelles variables jouent un rôle important dans celui-ci.

Nous avons donc essayé de créer un modèle simple qui effectue les mêmes prédictions que COMPAS, pour ensuite avoir accès aux différents poids de chaque variable. Pour effectuer cela, nous avons opté pour un modèle de régression logistique, avec une *accuracy* de 0.65, donc très similaire à celle de COMPAS. Nous avons alors pu observer le rôle prépondérent de l'âge dans notre modèle comme le montre la figure suivante.

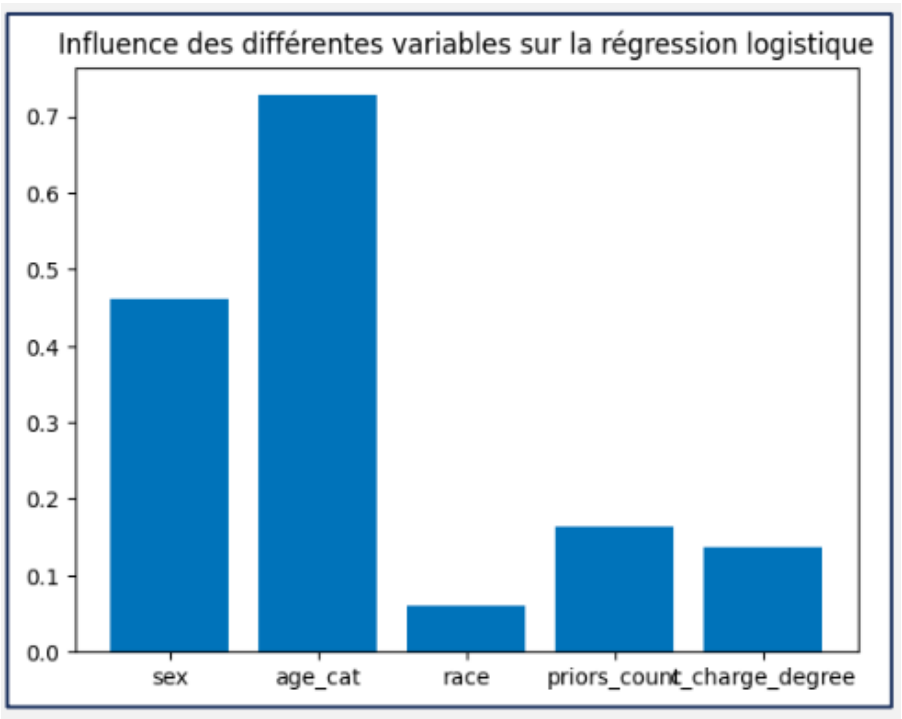

Figure 3 – Poids en valeur absolue attribué à chaque variable

Cela a encore confirmé notre première intuition sur l'importance de l'âge. En effet l'algorithme a beaucoup plus tendance à juger une personne jeune (< 25 ans) à risque par rapport au reste de la population. Mais cela n'explique pas encore le biais rencontré en fonction de l'ethnicté. C'est lorsque l'on s'intéresse à la répartition des jeunes en fonction de l'ethnicité que l'on peut comprendre une éventuelle explication à ce biais. La proportion de jeunes étant bien plus importante chez les Afro-Américains que chez les Caucasiens, (25 % de moins de 25 ans chez les Afro-Américains contre 16% chez les Caucasiens), cela pourrait être l'origine de cette disparité de jugement entre les deux ethnicités.

3.2 Analyse du biais par rapport à l'âge

Après avoir pris conscience de l'importance de l'âge dans notre propre modèle prédictif, il fallait s'assurer que c'était bien le cas dans COMPAS. Pour cela, nous avons retracé les courbes de distributions des scores en fonction de l'ethnicité en ne s'intéressant qu'à une seule catégorie d'âge à la fois. Nous avons pu remarquer plusieurs choses.

Premièrement, en regardant la catégorie des jeunes de moins de 25 ans, nous pouvons constater une distribution qui semble beaucoup plus équitable entre les deux ethnicités :

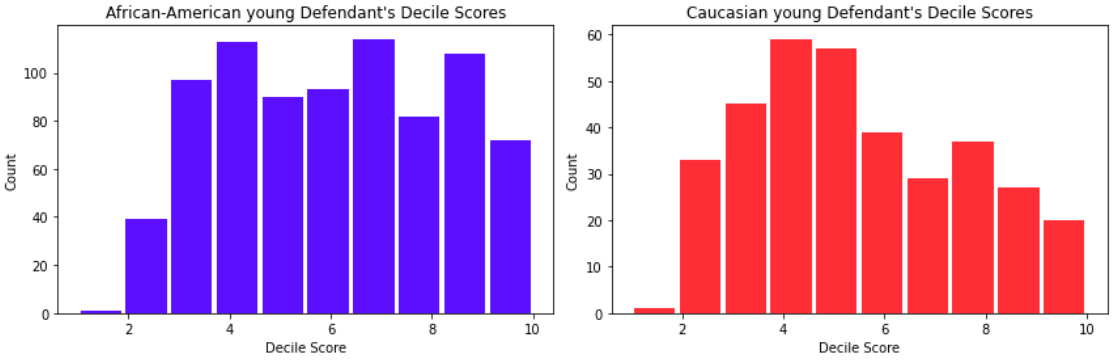

Figure 4 – Distribution des scores attribués aux jeunes de chaque ethnicité

Cependant, l'âge n'explique pas tout, comme on peut le voir sur le graphique de distribution du score chez les jeunes. Bien que le biais soit vraiment atténué par rapport aux courbes initiales, nous pouvons encore observer une différence non négligeable entre les deux catégories. Ensuite, pour les autres catégories d'âge, l'atténuation du biais est beaucoup moins flagrante, les Afro-Américains sont toujours jugés plus à risque que les Caucasiens et la distribution des scores suit la distribution globale.

En s'intéressant à la proportion de la population ayant récidivé dans les catégories d'âge de plus de 25 ans des deux ethnicités, nous pouvons voir que celle-ci est plus élevée chez les Afro-Américains que chez les Caucasiens. Cela pourrait être une piste de compréhension, l'algorithme COMPAS classe les Afro-Américains comme plus dangereux car les données sur lesquelles il a été entrainé montrent cela. Le fait que la proportion de récidivants soit plus élevée chez les Afro-Américains est peut-être une explication du biais mais en aucun cas elle ne le justifie. Pour toutes ces raisons, l'algorithme COMPAS ne semble pas très équitable entre les deux ethnicités. Nous avons donc essayé de créer un algorithme qui vérifie au mieux les critères de fairness mentionnés plus tôt, tout en gardant une *accuracy* au moins comparable.

4
RÉALISATION D'UN ALGORITHME PLUS PERFORMANT

4.1 Choix du modèle

Afin de commencer à créer notre propre algorithme pour prédire si une personne va récidiver ou non, il a fallu commencer par choisir un modèle. Nous en avons testé plusieurs (régression linéaire, régression logistique, random forest, SVM, XGBoost) et les avons d'abord comparés en fonction de leur accuracy (prédiction classique de nos données de test).

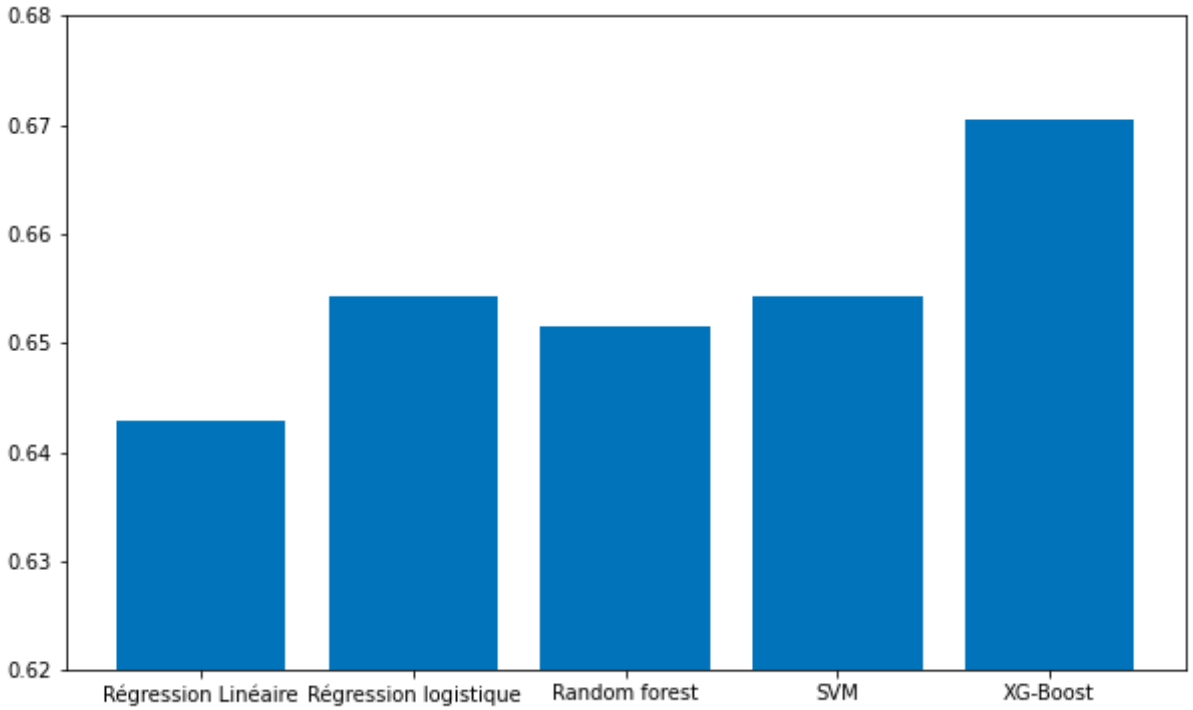

Figure 5 – Accuracy des différents modèles testés

Nous pouvons constater que sans grande surprise, le modèle XGBoost est le plus précis. Mais il est à noter que les différents modèles ont une accuracy très proches les uns des autres, et que par exemple nous avons une meilleure accuracy avec la régression logistique qu'avec une random forest, ce qui est assez surprenant.

Nous avons donc opté pour XGBoost et ce pour plusieurs raisons. Tout d'abord, dans le processus de sélection du modèle, ce choix de XGBoost s'avère être judicieux en raison de

sa précision initiale. En effet, avec une plus grande accuracy initiale, nous pourrions par la suite modifier certains paramètres pour débiaiser le modèle tout en gardant une très bonne précision. De plus, la flexibilité offerte par XGBoost en termes de réglage des hyperparamètres est un avantage crucial. XGBoost propose une large gamme d'hyperparamètres, permettant d'adapter précisément l'algorithme à notre ensemble de données spécifique. Cette capacité de personnalisation permet entre autre de prévenir l'overfitting et de maximiser les performances prédictives sur de nouvelles données.

En fin de compte, opter pour XGBoost comme modèle initial offre une base solide pour notre algorithme. L'opportunité de modifier les hyperparamètres permet de continuer à perfectionner le modèle tout en maintenant des performances élevées.

4.2 Démarche

Pour réaliser un algorithme plus performant, en utilisant XGBoost, nous nous sommes intéressés à l'accuracy et aux quatre critères de fairness (cf. 2.1) afin de les comparer aux résultats de l'algorithme de Northpointe (cf. 1.3). Nous voulions obtenir une accuracy similaire (de l'ordre de 65%) avec des critères de fairness mieux respectés. Nous nous sommes donc appuyés sur la littérature détaillée en 2.3, et avons testé différentes méthodes pour réduire le biais.

En nous basant sur les travaux de Besse et al. [7], nous avons utilisé des classifieurs différents pour les blancs et les Afro-Américains, obtenant des datasets d'environ 2500 (pour les blancs) et 3000 points (pour les Afro-Américains). Nous avons utilisé une cross-validation sur le dataset des blancs pour déterminer la taille optimale de l'échantillon de test. La figure suivante montre l'évolution de l'accuracy pour le modèle XGBoost.

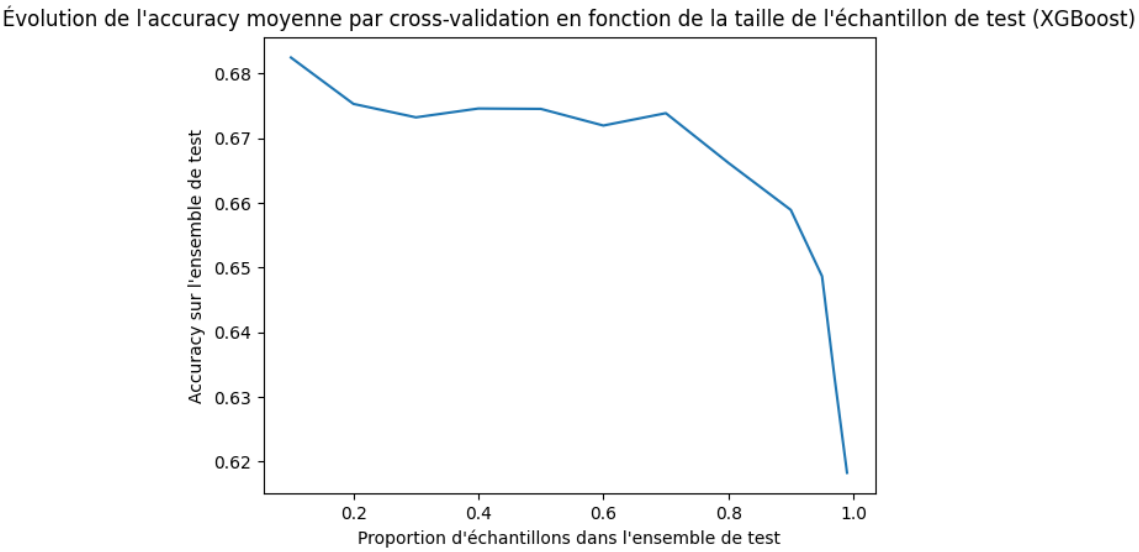

Figure 6 – Évolution de la précision en fonction de la taille de l'échantillon de test

On observe une baisse nette de l'accuracy lorsque l'ensemble de test est au-delà de 70% de la taille totale. On estime alors qu'une taille de 20% est optimale (5-fold), l'accuracy y étant la plus haute et semblant stabilisé.

Pour déterminer les incertitudes sur nos critères de performance et de fairness, nous avons utilisé le théorème de Student (cas où l'écart-type est inconnu) avec des variables aléatoires indicatrices. Nous obtenons alors des intervalles de confiance asymptotiques à 95% ($\alpha = 0.05$) :

$$\left[\bar{X}_n - t^{n-1}_{1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{X}_n + t^{n-1}_{1-\alpha/2}\frac{S}{\sqrt{n}}\right]$$

4.3 Méthodes utilisées avec XGBoost

Nous avons expérimenté différentes méthodes pour réduire le biais dans notre dataset, tout en conservant une accuracy similaire à celle de Northpointe.

Paliers Différents

En nous basant sur des travaux de Besse et al. [7], nous avons dans un premier temps utilisé des paliers différents de classification binaire pour les blancs (0.42) et les Afro-Américains (0.58), 0 correspondant à non-récidivant et 1 à récidivant. Ces choix de paliers résultent d'itérations successives, dans lesquelles nous avons cherché à obtenir un tradeoff entre les critères de fairness et l'accuracy. Il est possible de faire évoluer les paliers entre 0.4 et 0.6 sans observer une grande baisse d'*accuracy*, ainsi on obtient 2 paramètres à optimiser pour respecter les critères de fairness au mieux. Les résultats sont présentés dans le schéma suivant.

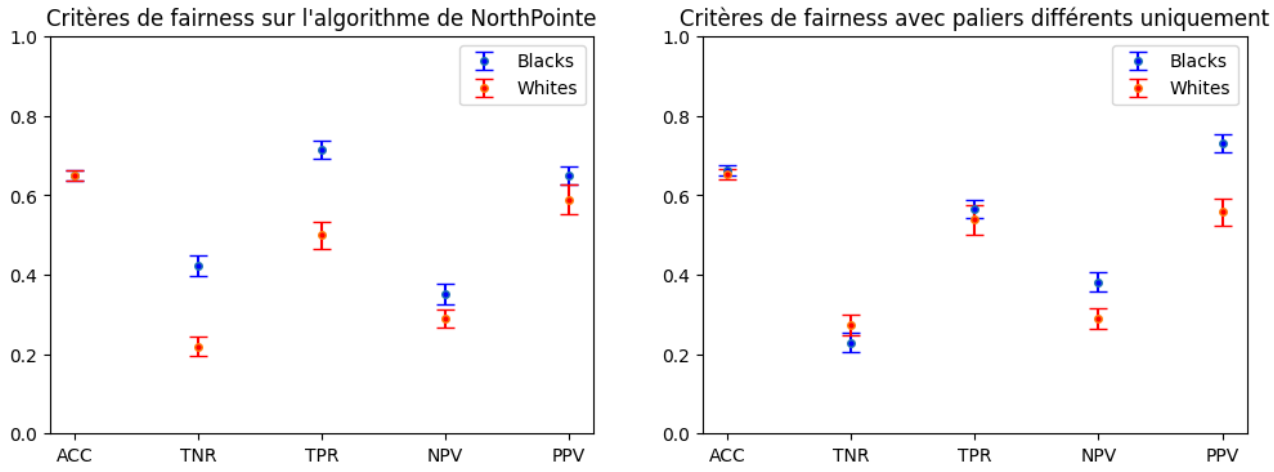

Figure 7 – Accuracy et de la fairness pour COMPAS et notre algorithme (paliers différenciés)

Les résultats nous montrent que le respect des critères TNR et TPR est possible. Cependant avec cette première méthode, il ne peut se faire qu'au prix d'une dégradation importante des critères NPV et PPV, ce qui est cohérent avec le théorème d'impossibilité vu en 2.2 [6].

Reweighting de la fonction de perte

L'importante hétérogénéité de notre dataset nous invite à nous pencher sur une méthode de reweighting de la fonction de perte, explorée par Li et al.[10] . En se fondant sur ce travail, notre objectif était de réaliser un reweighting du dataset de façon à maintenir un taux de récidive égal chez les Blancs et chez les Afro-Américains. Si ce reweighting est réalisé uniquement sur le dataset d'entraînement, alors on conserve le même ensemble de test et l'algorithme est donc parfaitement comparable avec l'algorithme de Northpointe.

Effectuer le reweighting sur l'ensemble de test est plus polémique. Une telle démarche utilise une hypothèse forte détaillée par Friedler et al. [5]. Nommée « *We are all equal* », cette hypothèse suppose qu'une une hétérogénéité du dataset initial doit nécessairement être corrigée car liée à des biais historiques ou culturels. Nous avons choisi de ne pas l'adopter, car c'est une hypothèse forte qui sort du cadre de la controverse entre ProPublica et Northpointe.

Nous avons donc effectué le reweighting uniquement sur l'ensemble d'entraînement (donc sur la fonction de perte). On observe alors, de façon intéressante, que les critères TNR et TPR sont naturellement respectés, tandis que NPV et PPV ne le sont pas.

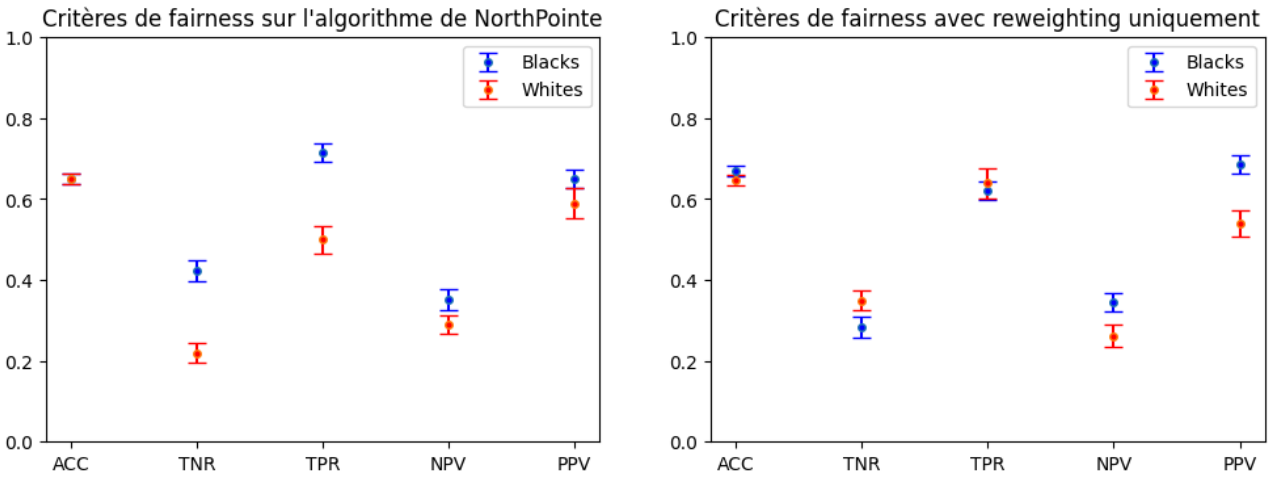

FIGURE 8 – Accuracy et fairness pour COMPAS et notre algorithme (avec reweighting)

En couplant cette méthode à la méthode des paliers, on obtient un algorithme optimisé, qui permet de mieux respecter les critères de quelques %.

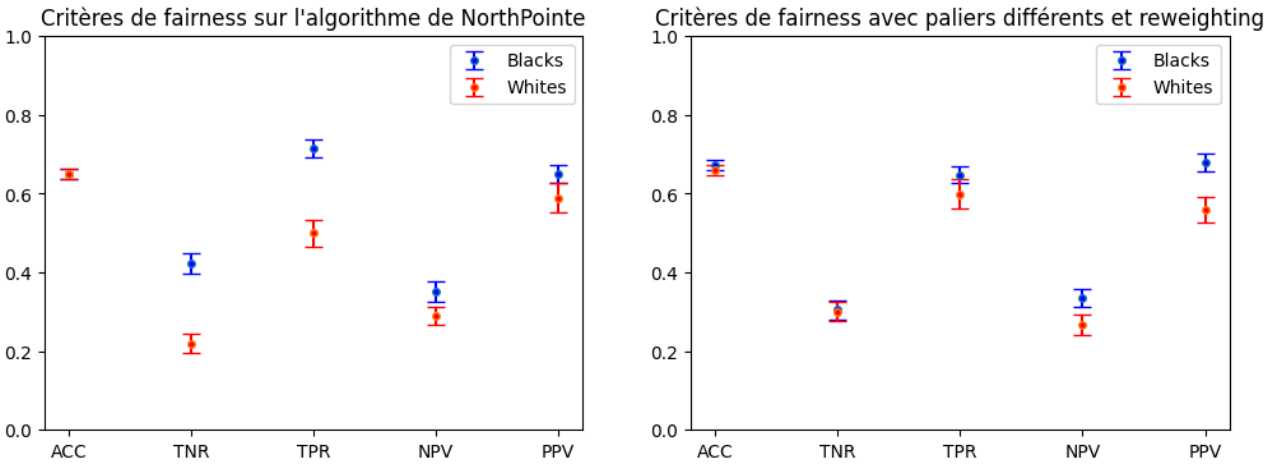

FIGURE 9 – Accuracy et fairness pour COMPAS et notre algorithme (reweighting et paliers différents)

4.4 ANALYSE DES RÉSULTATS, COMPARAISON AVEC COMPAS

Notre algorithme aboutit à des critères de fairness mieux respectés, avec un tradeoff plus équilibré. On représente ci-après les distributions de score de ce nouvel algorithme, plus équilibrées que celles de l'algorithme de NorthPointe.

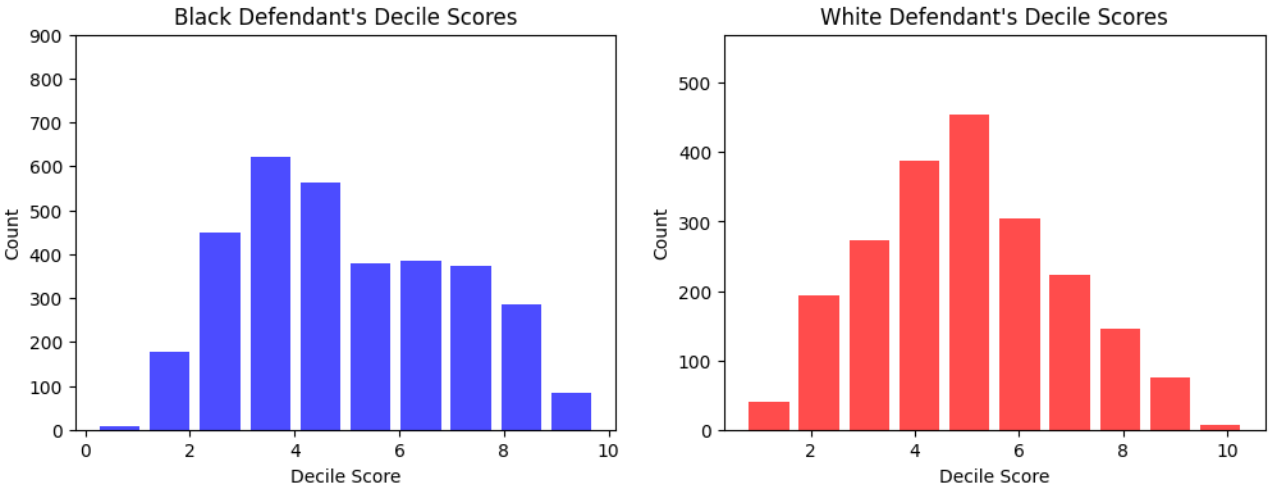

FIGURE 10 – Distributions de scores chez les Blancs et les Afro-Américains dans notre algorithme optimisé

Cependant, comparer notre algorithme à celui de NorthPointe semble vain tant qu'une juridiction claire n'est pas établie en matière de *fairness*. En partant de l'hypothèse que les quatre critères de *fairness* présentés se valent, nous avons souhaité établir un tradeoff plus équilibré.

Cependant, notre hypothèse n'est pas démontrable.

Il faut donc que la justice établisse une hiérarchie entre ces critères, et définisse comment les tradeoffs doivent être réalisés entre ces derniers. Le manque d'une telle juridiction est certainement lié à la jeunesse de la littérature en *fairness* : les articles de notre bibliographie ont pour la plupart été publiés entre 2020 et 2022. Un formalisme clair définissant la *fairness* manque, bien que Friedler et al. [5] représente un premier pas.

De plus, une juridiction est nécessaire pour définir les tradeoffs possibles entre *accuracy* et *fairness*. Aujourd'hui, la littérature omet souvent le critère d'*accuracy*, pourtant fondamental. Par exemple, un classifieur aléatoire aura des scores de TPR et TNR fair, mais l'*accuracy* ne sera évidemment pas acceptable. Une théorisation du tradeoff entre *accuracy* et *fairness* serait donc un pas important pour établir une telle juridiction.

5
RÉFLEXION ÉTHIQUE ET JURIDIQUE

La réflexion sur les implications éthiques et juridiques de ce type d'algorithme a été une grande composante de notre EA. En effet, nous avons réalisé deux plaidoiries (une en faveur de Northpointe et une en faveur de ProPublica) que nous avons soutenues devant une magistrate, ce qui nous a permis d'aller au-delà du cadre mathématique et d'ajouter une dimension pluridisciplinaire à notre travail.

À l'origine, les algorithmes comme COMPAS dans le domaine judiciaire ont été développés dans le but de faciliter la prise de décision, de réduire certains biais cognitifs potentiels (comme le « *hungry judge effect* »), ainsi que d'améliorer et d'accélérer le traitement des dossiers. Bien qu'ils ne soient pas destinés à se substituer entièrement à la décision humaine, ces outils sont conçus comme des aides à la synthèse d'informations pour les juges et donc à la prise de décision. Cependant, malgré leur objectif louable, l'utilisation de tels algorithmes suscite des préoccupations fondamentales.

Tout d'abord, ce type d'algorithmes peut amplifier des inégalités déjà présentes. Les données historiques utilisées pour entraîner ces algorithmes peuvent refléter des discriminations systémiques existantes, introduisant ainsi des préjugés dans les prédictions futures. Cela perpétue et amplifie les inégalités présentes dans le système judiciaire, ce qui semble être le cas avec COMPAS et les Afro-Américains.

Par ailleurs, l'opacité qui entoure les mécanismes internes de ces algorithmes, détenus par des entreprises privées, est particulièrement problématique car elle crée un voile de mystère autour de décisions qui affectent directement la vie des individus. Cette absence de transparence compromet le principe fondamental de la justice selon lequel les personnes devraient comprendre les motifs et les critères qui influent sur les décisions qui les concernent. Ainsi, l'utilisation de tels algorithmes peut compromettre le droit à un procès équitable. Les individus sont confrontés à des décisions algorithmiques souvent complexes sans avoir une compréhension adéquate de la logique derrière ces choix. Cela peut entraver leur capacité à contester efficacement les résultats et à faire valoir leurs droits devant la justice.

Ainsi, l'utilisation de tels algorithmes dans le système judiciaire présente des risques sérieux en termes de transparence, de perpétuation des inégalités et de respect des droits individuels. Cependant, il semble que le risque le plus grave ne soit pas l'utilisation en soi d'algorithmes, et ce même dans des domaines sensibles, mais le manque de transparence de ces algorithmes s'ils sont détenus par des opérateurs privés. C'est ici que la question juridique intervient avec la régulation de ces pratiques.

En France, l'utilisation de tels algorithmes est autorisée, mais encadrée par des instruments

juridiques tels que la charte éthique d'utilisation de l'intelligence artificielle dans les systèmes judiciaires, adoptée par le Conseil de l'Europe le 8 décembre 2018. Cette charte énonce des principes clés, parmi lesquels figurent la transparence et la maîtrise par l'utilisateur, soulignant l'importance de rendre explicites les processus décisionnels automatisés et de garantir une participation significative des individus. De plus, la législation française, notamment la loi informatique et libertés, interdit les décisions juridiques entièrement automatisées. Ces dispositions sont conformes aux principes du RGPD (Règlement Général à la Protection des Données personnelles), qui mettent l'accent sur la transparence, le droit à l'information et la possibilité pour les individus de contester les décisions automatisées les concernant. Ainsi, en France, l'utilisation de l'intelligence artificielle est soumise à un cadre juridique rigoureux visant à protéger les droits fondamentaux des individus, notamment dans des domaines sensibles comme le domaine judiciaire.

CONCLUSION

L'étude approfondie de l'algorithme COMPAS et de ses implications dans le système judiciaire a mis en lumière des enjeux cruciaux aux intersections des mathématiques appliquées, de l'éthique et du droit. Cet EA a été l'occasion de découvrir la question des biais dans les algorithmes et d'étudier la littérature en matière de *fairness*. Au-delà de l'étude de l'état de l'art, notre travail a comporté 3 volets.

- **L'analyse des données** afin d'ajouter des éléments supplémentaires à la controverse. L'influence significative de l'âge sur les prédictions a ainsi été identifiée, soulignant la complexité des facteurs contribuant aux résultats de l'algorithme, et la difficulté de quantifier les biais par rapport à une variable sensible.
- **La création d'un nouvel algorithme** visant à améliorer la performance tout en atténuant les biais observés. L'adoption du modèle XGBoost s'est révélée judicieuse en raison de sa précision et de sa flexibilité, permettant des ajustements ultérieurs. Différentes approches, telles que l'utilisation de paliers de classification distincts et le reweighting de la fonction de perte, ont été explorées pour améliorer la justice algorithmique. L'analyse des résultats, comparée à COMPAS, a révélé une réduction des biais avec un tradeoff plus équilibré entre la précision et les critères de *fairness*.
- **Une réflexion éthique et juridique.** Au cours de notre EA, la simulation du procès Northpointe-ProPublica devant une magistrate a motivé une réflexion éthique et juridique approfondie. Les implications éthiques soulignent le risque d'amplification des inégalités existantes et le compromis du droit à un procès équitable. L'opacité entourant les mécanismes internes de ces algorithmes détenus par des entreprises privées est également problématique, soulignant le besoin urgent d'une régulation claire.

En conclusion, la question centrale réside dans l'équilibre délicat entre la recherche de l'efficacité algorithmique et le respect des principes éthiques et juridiques fondamentaux. La nécessité d'une juridiction claire, définissant la *fairness*, devient impérative pour guider le développement et l'utilisation responsables de ces technologies dans le domaine judiciaire. La transparence et la régulation sont les piliers essentiels pour garantir que l'IA, lorsqu'elle est utilisée dans des contextes sensibles, respecte les droits fondamentaux et ne perpétue pas les inégalités existantes.

RÉFÉRENCES

[1] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

[2] Christina Mendoza William Dieterich and Tim Brennan. Compas risk scales : Demonstrating accuracy equity and predictive parity. *Northpointe Inc.*, 2016.

[3] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini1. A clarification of the nuances in the fairness metrics landscape. *Nature*, 2022.

[4] Sahil Verma and Julia Rubin. Fairness definitions explained. *ACM/IEEE International Workshop on Software Fairness*, 2018.

[5] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness. *Proceedings of The Algorithmic Fairness through the Lens of Causality and Robustness, PMLR*, 2022.

[6] Pratyush Garg, Carlos Scheidegger, and Suresh Venkatasubramanian. Fairness metrics : A comparative analysis. *Proceedings of The Algorithmic Fairness through the Lens of Causality and Robustness, PMLR*, 2020.

[7] P. Besse, E. del Barrio, P. Gordaliza, J-M. Loubes, and L. Risser. A survey of bias in machine learning through the prism of statistical parity for the adult data set. 2020.

[8] Marco Scutari, Francesca Panero, and Manuel Proissl. Achieving fairness with a simple ridge penalty. 2020.

[9] Manisha Padala and Sujit Gujar. Fnnc : Achieving fairness through neural networks. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, 2020.

[10] Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. *Proceedings of Machine Learning Research*, 2022.