Discriminations intersectionnelles : approfondir l'évaluation de l'équité algorithmique. L'exemple de la prédiction de la réussite académique avec des données issues de cours en ligne

Intersectionalinity : deepen algorithmic fairness evaluation. The case study of academic performance prediction using data from online courses

Mélina VERGER[1]; François BOUCHET[1]; Sébastien LALLÉ[1]; Vanda LUENGO[1]

[1] *Sorbonne Université, CNRS, LIP6, F-75005 Paris, France*

Résumé. Evaluer l'équité algorithmique des modèles prédictifs utilisés en éducation est devenu crucial pour garantir que, déployés, ils ne soient pas biaisés envers certains apprenants. Jusqu'à présent, les analyses se sont concentrées sur l'évaluation de l'équité algorithmique vis-à-vis d'attributs sensibles, comme le genre, présents dans les données, indépendamment les uns des autres. Or, la théorie de l'intersectionnalité de Crenshaw (1989) défend l'idée que les influences conjointes de plusieurs attributs sensibles produisent des discriminations uniques et différentes pour certaines sous-groupes d'individus. Ainsi, nous proposons dans cet article d'étendre les travaux de Verger, Bouchet *et al.* (2023) avec des analyses supplémentaires sur les discriminations intersectionnelles présentes dans les prédictions des modèles. Ces modèles ont été utilisés dans le cadre de la prédiction de la réussite à des cours en ligne, au moyen de données éducatives ouvertes, plus précisément les données du corpus OULAD (Kuzilek *et al.*, 2017). Nos résultats ont permis de mettre en lumière des discriminations algorithmiques qui n'étaient pas identifiables à partir des analyses classiques ainsi que de déterminer l'influence de chaque attribut sur les discriminations grâce à leurs interactions avec les autres attributs.
Mots-clés : équité algorithmique, intersectionnalité, prédiction

Abstract. *Assessing the algorithmic fairness of predictive models used in education has become crucial to ensure that, when deployed, they are not biased in favor or against certain learners. Until now, analyses have focused on assessing algorithmic fairness with regard to sensitive attributes in the data, such as gender, independently of each other. However, Crenshaw (1989)'s theory of intersectionality defends the idea that the influences of several sensitive attributes together produce unique and different discriminations for certain sub-groups of individuals. Thus, in this paper, we propose to extend (Verger, Bouchet et al., 2023)'s work with additional analyses of intersectional discriminations that are in the outcomes of predictive models. These models were used in the context of predicting success in online courses, using open educational data, specifically data from OULAD (Kuzilek et al., 2017). Our results shed light on algorithmic discriminations that were not identifiable from traditional analyses, as well as determining the influence of each attribute on discriminations through their interactions with other attributes.*
Keywords : *algorithmic fairness, intersectionality, prediction*

1. INTRODUCTION

Dans cette section, nous examinons pourquoi et comment nous nous intéressons aux discriminations intersectionnelles. Nous présentons d'abord le vocabulaire qui sera employé tout au long de l'article (Section 1.1). Puis, nous expliquons le contexte de notre recherche (Section 1.2), suivi de ses motivations et questions de recherche (Section 1.3). Enfin, nous décrivons les contributions attendues (méthodes et résultats principaux) (Section 1.4) ainsi que le plan de l'article (Section 1.5).

1.1. TERMINOLOGIE

Avant d'entrer dans les détails, nous précisons quelques éléments terminologiques. Premièrement, l'*équité algorithmique*, qui est en fait une traduction des termes anglais plus largement utilisés *algorithmic fairness*, fait référence non pas à une équité mais à une égalité de traitement des individus par les systèmes composés d'au moins un modèle d'apprentissage automatique. L'égalité de traitement doit être faite, indépendamment de caractéristiques telles que le sexe ou l'origine ethnique [1] (Barocas *et al.*, 2019). Par exemple, un modèle qui prédit l'admission à l'université doit produire des résultats indépendants du genre de l'apprenant. À l'inverse, nous parlerons de *discriminations algorithmiques* lorsqu'un système s'avèrera ne pas produire cette égalité de traitement en fonction de l'une ou l'autre des caractéristiques en question. Dans le contexte de cet article, *équité* et *discrimination* seront parfois utilisés sans l'adjectif *algorithmique* par soucis de simplicité.

Deuxièmement, ces caractéristiques, auxquelles nous ne voulons pas que les résultats d'un système accordent un traitement discriminatoire, sont appelées *attributs sensibles* (ou en anglais *protected attributes* ou *sensitive features* (Kizilcec et Lee, 2022 ; Pessach et Shmueli, 2022)). Ces attributs peuvent correspondre à des données dites sensibles en tant que données personnelles, telles que des données socio-démographiques (comme les deux évoquées précédemment), mais nous emploierons ces termes d'*attributs sensibles* de manière plus générale pour désigner les attributs vis-à-vis desquels nous ferons des analyses d'équité.

1.2. CONTEXTE

L'équité algorithmique, c'est-à-dire l'équité dans les systèmes utilisant de l'apprentissage automatique, combinant des modèles mathématiques avec des données, tels que les algorithmes de prédictions, a acquis une importance cruciale en raison de l'utilisation croissante de ces systèmes informatiques. Un exemple d'un tel système en éducation est celui des systèmes adaptatifs, qui utilisent des algorithmes de prédiction pour prédire si l'élève va réussir ou non une tâche, un exercice, un cours, un module, ce qui permet, entre autres, d'adapter l'exercice, la tâche, la recommandation de ressources, le cours ou le parcours.

De la même manière qu'il est courant de tester plusieurs algorithmes de prédiction sur un même corpus de données, car ne donnant pas les mêmes résultats, il est également possible que ces algorithmes donnent des résultats différents vis-à-vis des groupes d'individus. Aussi, l'algorithme le plus performant en termes de prédiction n'est pas nécessairement l'algorithme le plus *équitable* au regard de différents groupes d'individus. Si l'évaluation de la performance prédictive est classique dans ce type de systèmes, l'évaluation de l'équité algorithmique est quant à elle plus récente et est devenue nécessaire afin de garantir que les

1. Dans l'Union Européenne (UE), les analyses sur l'origine ethnique ne peuvent être conduites du fait du Règlement Général sur la Protection des Données (RGPD) interdisant la collecte de ce type d'information.

systèmes déployés ne soient pas biaisés en faveur ou au détriment de certains groupes d'individus (Buolamwini et Gebru, 2018 ; Gardner *et al*., 2019 ; Lee et Kizilcec, 2020 ; Mehrabi *et al*., 2022 ; Verger, 2022). Ainsi, l'équité algorithmique constitue un enjeu éthique, mais aussi un impératif sociétal, émanant notamment de directives et réglementations de plusieurs instances [2].

Les discriminations algorithmiques peuvent apparaître à différentes phases de développement des systèmes informatiques, passant par la collecte des données, leur pré-traitement, l'entraînement du modèle et enfin leur usage (Verger, 2022). Les causes de ces discriminations se retrouvent donc soit dans les données, les algorithmes ou les interventions humaines suite aux résultats. En ce qui concerne les sources de discrimination liées aux données, elles peuvent être dues aux discriminations dites réelles ou historiques (c'est-à-dire observables depuis longtemps dans une société, et se reflétant dans les données), ou bien elles peuvent être relatives à une représentation, une collecte ou un calibrage inadéquates des données. C'est donc dans la phase de collecte et pré-traitement que ces discriminations peuvent être étudiées. Ici, nous cherchons à évaluer l'équité algorithmique en sortie d'entraînement du modèle, avant son usage. Autrement dit, nous allons analyser les discriminations algorithmiques présentes dans les résulats du modèle.

1.3. Motivations et questions de recherche

Dans un travail antérieur (Verger, Bouchet *et al*., 2023), nous avions évalué l'équité algorithmique selon certains attributs sensibles (voir Section 1.1), étudiés individuellement. Nous avions notamment combiné les données ouvertes OULAD (Kuzilek *et al*., 2017) avec des modèles prédictifs très couramment utilisés en éducation (régression logistique, arbre de décision, k plus proches voisins et naïf bayésien). Nous les avions évalués avec une nouvelle mesure d'équité, la MADD (*Model Absolute Density Distance*), que nous avions spécifiquement proposée dans Verger, Lallé *et al*. (2023) pour éviter un écueil courant dans l'évaluation de l'équité algorithmique (voir Section 2). Pour résumer, cette évaluation a donné lieu à une étude numérique et visuelle de l'influence de chaque attribut sensible sur l'équité des résultats des modèles. Ceci nous avait permis de montrer qu'il n'y avait pas de lien direct entre les discriminations observées dans les données et les discriminations obtenues en sortie des modèles. Ainsi, malgré le biais de genre dans les données du cours de sciences sociales (fortes corrélations avec l'attribut genre, forts déséquilibres entre les deux groupes le constituant avec une large majorité de femmes), notre analyse montre que c'est un autre attribut sensible, "pauvreté", qui est a l'origine des discriminations algorithmiques les plus importantes. Cela montre l'importance d'évaluer en profondeur l'équité des modèles produits par ce type d'algorithme en plus des évaluations liées aux autres sources de discrimination.

Ainsi, dans cet article, nous nous proposons d'aller plus loin dans l'évaluation de l'équité algorithmique en considérant non pas l'influence individuelle, mais l'influence simultanée de plusieurs attributs sensibles. En effet, la théorie de l'intersectionnalité de Crenshaw (1989) défend l'idée que les influences de plusieurs attributs ensemble produisent des discriminations uniques et différentes pour certains sous-groupes d'individus. Un exemple souvent rattaché à cette théorie est celui des discriminations expérimentées en tant que femmes noires, qui diffèrent des discriminations entre hommes et femmes ou entre personnes noires et blanches respectivement (Buolamwini et Gebru, 2018). Par conséquent, nous allons examiner les *discriminations intersectionnelles*, c'est-à-dire les discriminations issues de l'in-

2. Par ordre chronologique : Règlement Général sur la Protection des Données (2016) au niveau européen, *California Consumer Privacy Act* (2018) au niveau des États-Unis, Principes de l'OCDE (Organisation de coopération et de développement économiques) sur l'intelligence artificielle (2019) au niveau international, et prochainement l'*Artificial Intelligence Act* au niveau européen.

fluence de plusieurs attributs sensibles en même temps.

Deux questions de recherche se dégagent alors :

QR1 : Comment évaluer l'influence de plusieurs attributs sensibles simultanément ?

QR2 : Découvre t-on des discriminations algorithmiques supplémentaires quand on considère les individus à l'intersection de plusieurs attributs sensibles ?

1.4. CONTRIBUTIONS

Pour répondre à ces questions, nous suivons une démarche exploratoire et nous proposons de poursuivre les analyses de Verger, Bouchet *et al*. (2023) en évaluant avec la même mesure, MADD, l'équité vis-à-vis des quatre attributs sensibles considérés précédemment mais pris conjointement : le genre, l'âge, le niveau de pauvreté et le handicap. Pour cela, nous nous intéresserons à tous les *groupes intersectionnels* possibles formés par ces quatre attributs (voir Section 6.2). Plus précisément, un *groupe intersectionnel* est défini comme étant un groupe à l'intersection de plusieurs attributs (Gohar et Cheng, 2023), comme par exemple le groupe des « jeunes hommes aisés sans handicap » qui est à l'intersection des attributs âge, genre, niveau de pauvreté et handicap (plus de détails en Section 4).

Les résultats de cet article sont doubles. D'une part, nous proposons une approche d'évaluation pour l'analyse des discriminations intersectionnelles (QR1) et, d'autre part, les analyses permettent de mettre en lumière des discriminations qu'on ne peut voir avec l'analyse des attributs individuels seule et de comprendre de manière plus fine l'influence de chaque attribut sur les discriminations grâce à leurs interactions avec les autres attributs (QR2).

Cet article participe à faire avancer la recherche sur les discriminations algorithmiques et leur évaluation dans les modèles prédictifs en éducation. Comme indiqué plus haut, les discriminations algorithmiques pouvant être ou non différentes des discriminations réelles, ce type d'analyses permet de fournir des éclairages sur les implications de l'utilisation de ces modèles ainsi que sur les discriminations qui pourraient être découvertes. À notre connaissance, cette approche intersectionnelle pour l'évaluation de l'équité algorithmique n'a été faite qu'une fois avec des données éducatives par Zambrano *et al*. (2024), mais sans parvenir à établir de discriminations entre groupes intersectionnels. Notre étude est ainsi la première en éducation à montrer l'utilité de l'approche intersectionnelle pour mettre en lumière des discriminations algorithmiques qui n'auraient pas pu être détectées autrement. Par ailleurs, cette approche est particulièrement destinée aux chercheuses, chercheurs, développeuses et développeurs de modèles prédictifs en éducation et nous mettons à disposition les données, le code documenté [3] ainsi que la librairie Python `maddlib` [4] que nous avons développée pour faciliter ces analyses.

1.5. PLAN DE L'ARTICLE

L'article est organisé comme suit. La Section 2 décrit l'état de l'art de l'évaluation de l'équité algorithmique en éducation spécifiquement pour la tâche de prédiction de la réussite au niveau des cours. La Section 3 présente la mesure d'équité MADD, et la Section 4 comment analyser l'équité avec des groupes intersectionnels. Puis, la Section 5 introduit le cadre expérimental des analyses réalisées en Section 6. Enfin, la Section 7 examine les forces et limites des expériences et des méthodes proposées, avant de conclure en Section 8.

3. https://github.com/melinaverger/MADD
4. https://pypi.org/project/maddlib

2. ÉTAT DE L'ART

La problématique de cet article portant sur l'évaluation de l'équité algorithmique en considérant l'intersectionnalité, avec comme cas d'étude les systèmes de prédiction de réussite dans des cours en ligne, nous allons d'abord introduire rapidement la notion de réussite à des cours en ligne et les méthodes informatiques de prédiction utilisées (Section 2.1), pour ensuite présenter les travaux sur l'évaluation de l'équité dans ce type de systèmes (Section 2.2), pour finir avec la prise en compte de l'intersectionnalité (Section 2.3).

2.1. LA PRÉDICTION DE LA RÉUSSITE À DES COURS

De manière générale, la capacité à prédire les performances des apprenantes et apprenants dans un cours, ou un parcours, peut permettre d'améliorer les résultats de l'enseignement (Hellas *et al.*, 2018). Comme l'indiquent ces auteurs, des recherches en psychologie et en sciences de l'éducation cherchent à comprendre les facteurs de réussite dite académique depuis au moins un siècle. Les travaux utilisant des algorithmes de prédiction sont quand à eux apparus plus tard, avec l'accès aux données académiques.

Dans cet article, nous nous intéressons aux recherches sur la prédiction de la réussite en s'appuyant sur des données académiques. Ces recherches se concentrent principalement sur la prédiction de la performance à des cours, notamment en termes de réussite ou d'échec, ainsi que sur leur persévérance ou leur abandon dans des contextes d'apprentissage en ligne ou hybride. Comme l'indique la revue systématique d'Hellas *et al*. (2018), la recherche dans ce domaine étudie les caractéristiques (attributs ou facteurs) susceptibles d'être exploitées pour prédire la réussite ainsi que les algorithmes permettant d'améliorer ces prédictions.

Nous pouvons constater que la prédiction de la réussite à partir des données de cours est actuellement une recherche active dans les domaines des learning analytics, fouille de données éducatives et intelligence artificielle pour l'éducation (Romero et Ventura, 2020). En effet, elle constitue, avec la prédiction précoce d'abandon notamment, une des tâches couramment réalisées dans le domaine de l'analyse des données éducatives. C'est pourquoi l'équité algorithmique est étudiée dans le cadre de cette tâche de prédiction (Deho *et al.*, 2022). Bien que ces recherches puissent être réductrices quant aux facteurs influençant la réussite et l'échec, le centre de notre travail ici est l'évaluation de l'équité algorithmique.

Du point de vue des techniques informatiques, d'après différentes revues de littérature (Hellas *et al.*, 2018 ; Korkmaz et Correia, 2019), prédire la réussite à des cours est le plus souvent représenté par un problème de classification binaire (réussite/échec). Plusieurs types de modèles sont couramment utilisés, tels que les réseaux bayésiens, les arbres de décision, ou la méthode des k plus proches voisins. Dans la continuité de ces pratiques, nous expérimentons notre approche avec quatre types de modèles de classification binaire très courants en éducation (Hellas *et al.*, 2018 ; Korkmaz et Correia, 2019) et en particulier sur le corpus OULAD (Alhakbani et Alnassar, 2022 ; Kuzilek *et al.*, 2017). Nous détaillerons ces choix dans la Section 5 dédiée aux expériences. Par ailleurs, le corpus OULAD a été utilisé dans plusieurs travaux de prédiction connexes (réussite/échec, abandon/complétion des cours, etc. voir (Alhakbani et Alnassar, 2022)), mais sans analyse des discriminations algorithmiques jusqu'à ce que nous les initions dans Verger, Bouchet *et al*. (2023).

2.2. ÉVALUER L'ÉQUITÉ ALGORITHMIQUE

Pour évaluer l'équité algorithmique, il existent trois approches : causale, de similarité et statistique. Pour la première famille, qui traite de l'équité à l'aide de l'approche causale, l'absence d'inéquité est considérée lorsqu'on change l'appartenance d'un individu à un groupe pour un autre (par exemple "homme" pour "femme") et que son résultat reste inchangé

(Kilbertus *et al*., 2017). Dans l'approche de similarité, l'absence d'inéquité est considérée lorsque deux individus similaires ont obtenu des résultats similaires. Les techniques de cette approche diffèrent selon la distance qu'elles emploient pour mesurer cette notion de similarité. Pour la troisième famille, fondée sur les statistiques, l'absence d'inéquité est considérée lorsqu'un système produit des résultats similaires entre les groupes. Contrairement aux techniques fondées sur la similarité, celles-ci ne se concentrent pas sur des paires d'individus mais sur des groupes, d'où leur autre appellation : « équité de groupe ». Ces techniques consistent donc à quantifier les différences de performance d'un modèle entre les groupes, par exemple grâce à des comparaisons du taux de d'erreur.

Les trois approches ont leurs avantages et inconvénients. Cependant, il s'avère plus difficile d'utiliser l'approche causale car elle requiert de construire des graphes causaux, ce qui est difficile en pratique (Pearl, 2009). De la même façon, l'approche de similarité nécessite une mesure pertinente qui permette de comparer un à un des individus, ce qui est également difficile à mettre en place sans produire d'autres biais. C'est pour cette raison que l'approche statistique est la plus utilisée. Comme évoqué juste avant, pour évaluer l'équité algorithmique avec l'approche statistique, il existe des techniques différentes. Elles peuvent se regrouper en trois familles distinctes : indépendance, séparation et suffisance. L'objectif des mesures d'indépendance est d'évaluer si les résultats sont indépendants de l'appartenance à un groupe. L'objectif des mesures de séparation est d'évaluer si les résultats sont indépendants de l'appartenance à un groupe, mais conditionnés par les valeurs cibles (ou étiquettes, *labels*) dans les données. L'objectif des mesures de suffisance est d'évaluer si les valeurs cibles sont indépendantes de l'appartenance à un groupe en fonction des résultats. Tous ces types de mesures statistiques sont des moyens valables mais distincts d'interpréter l'équité entre les groupes. Ils présentent tous également des avantages et des limites, et le type de mesure ainsi que la manière d'utiliser les valeurs cibles dans l'évaluation dépendent du contexte. Pour avoir plus des détails sur ces approches, nous conseillons de consulter Castelnovo *et al*. (2022).

Dans le cadre de la prédiction de la réussite académique, quelques travaux se sont concentrés sur l'évaluation de l'équité algorithmique par des approches statistiques. C'est le cas notamment de Gardner *et al*. (2019), Hu et Rangwala (2020), et Lee et Kizilcec (2020), dont les études, menées exclusivement aux Etats-Unis, ont concerné également des modèles de classification binaire pour cette tâche.

Ces travaux ont utilisé différentes mesures d'équité (Caton et Haas, 2020 ; Verma et Rubin, 2018), comme l'égalité de la précision des prédictions d'un modèle entre les différents groupes (indépendance) ou l'égalité du taux d'erreur de prédiction (séparation) (Baker et Hawn, 2022). Un modèle était alors considéré équitable s'il produisait des performances prédicitives similaires entre tous les groupes. Cependant, évaluer l'équité algorithmique à partir des performances prédictives des modèles n'est pas synonyme d'absence de discrimination algorithmique. En effet, un modèle peut produire des erreurs en même quantité pour différents groupes mais elles peuvent être plus sévères pour un groupe que pour un autre car plus éloignées de la réalité et donc avec des implications plus négatives. C'est pourquoi, et en réponse à l'écueil mentionné en Section 1, nous avions développé une nouvelle mesure, la MADD (Verger, Lallé *et al*., 2023), indépendante de la performance prédictive, que nous utilisons à nouveau ici pour conduire des analyses d'équité algorithmique supplémentaires à celles de Verger, Bouchet *et al*. (2023). Cette mesure sera présentée en Section 3.

De plus, ce sont souvent les mêmes attributs sensibles qui sont étudiés, à savoir le genre et l'origine ethnique (Baker et Hawn, 2022). En effet, Gardner *et al*. (2019) ont comparé les performances prédictives des modèles utilisés par rapport au genre, et Hu et Rangwala (2020) ou Lee et Kizilcec (2020) par rapport au genre et à l'origine ethnique des apprenants. Bien qu'ils soient pertinents à étudier, Baker et Hawn (2022) appellent à considérer une

plus grande diversité d'attributs sensibles pour obtenir davantage d'informations sur leur influence dans l'équité algorithmique. Par conséquent, nos analyses effectuées avec le corpus OULAD (Kuzilek *et al*., 2017), à la fois dans ce présent article et dans Verger, Bouchet *et al*. (2023), nous permettent d'étudier quatre attributs sensibles : le genre, l'âge, le niveau de pauvreté et le handicap ; les trois derniers ayant été très rarement considérés dans le contexte des analyses d'équité algorithmique en éducation.

2.3. Équité et discriminations intersectionnelles

Jusqu'à très récemment, toutes les études d'équité algorithmique en éducation proposaient une évaluation attribut sensible par attribut sensible (Baker et Hawn, 2022). Par exemple, le score d'équité du genre était calculé, puis celui de l'origine ethnique, indépendemment. Si cette approche d'évaluation individuelle permet bien de classer les attributs selon leur score obtenu avec n'importe quelle mesure d'équité algorithmique, elle suppose que ces attributs soient indépendants et que les discriminations le soient également. Or, comme mentionné en introduction (Section 1), d'après la théorie de l'intersectionnalité de Crenshaw (1989), les influences conjointes de plusieurs attributs produisent des discriminations uniques et différentes pour certains groupes intersectionnels. Par exemple, les discriminations expérimentées en tant que femmes noires diffèrent des discriminations soit entre hommes et femmes soit entre personnes noires et blanches respectivement (Buolamwini et Gebru, 2018 ; Evans-Winters, 2021). Ainsi, au lieu de calculer un score pour le genre et pour l'origine ethnique séparément, on voudrait un score qui prenne en compte, par exemple, le fait d'être à la fois une femme et d'une certaine origine ethnique.

À notre connaissance, il n'existe qu'une seule étude d'équité algorithmique en éducation qui présente une analyse au niveau de groupes intersectionnels (Zambrano *et al*., 2024). Dans cette étude, l'équité de deux modèles prédictifs est étudiée pour différents attributs démographiques, incluant l'origine ethnique, le handicap, le genre, la langue maternelle et le niveau de pauvreté, sur une population d'apprenants de plusieurs lycées d'une petite ville du Nord-Est des États-Unis. L'analyse de l'équité est faite en calculant la performance des modèles prédictifs pour tous les groupes possibles à l'intersection entre l'origine ethnique d'un côté, et les autres attributs de l'autre, par exemple les « hispaniques handicapés » ou les « métisses pauvres ». De cette manière, les auteurs identifient les sous-groupes pour lesquels les performances des modèles sont sensiblement plus mauvaises que pour les autres. Leurs résultats ne permettent cependant pas d'identifier de discriminations pour aucun des attributs ni aucune de leurs intersections, peut-être en raison de la taille restreinte de leur base de données (5 000 lycéennes et lycéens provenant d'une même ville). Dans notre étude, nous proposons, comme Zambrano *et al*. (2024), une analyse des discriminations algorithmiques au niveau des attributs seuls (voir Section 6.1) et une autre au niveau des groupes intersectionnels (voir Section 6.2). Nous utiliserons pour cela le corpus OULAD qui est plus large que le corpus utilisé par Zambrano *et al*. (2024) et pour lequel nous avions déjà montré la présence de discriminations algorithmiques pour plusieurs attributs pris séparément. Ce corpus présente donc un potentiel intéressant pour l'étude des groupes intersectionnels, et nos résultats mettent justement en lumière l'utilité de l'approche intersectionnelle sur ce corpus. De plus, nous utiliserons une métrique d'équité algorithmique (la MADD, comme évoqué précédemment en Section 2.2), plutôt que de simplement comparer les performances prédictives des modèles comme dans l'étude de Zambrano *et al*. (2024), afin de pouvoir quantifier les possibles discriminations entre plusieurs groupes intersectionnels.

3. LA MADD, MESURE D'ÉQUITÉ ALGORITHMIQUE

A présent, nous présentons la mesure d'équité algorithmique MADD [5], utilisée à la fois pour les analyses selon les attributs sensibles individuels en Section 6.1, et pour les analyses selon les groupes intersectionnels réalisées en Section 6.2. En complément des explications présentes dans cette Section 3, nous informons les lectrices et lecteurs que l'article de Verger, Bouchet *et al*. (2023) décrit de manière additionnelle comment exploiter l'aspect visuel de la MADD pour des analyses d'équité, ce dont nous ne nous servirons pas ici car nous chercherons à quantifier les discriminations.

3.1. PRÉLIMINAIRES

Considérons un modèle de classification binaire $\mathcal{C}$ pour la prédiction de la réussite à un cours en ligne. Pour pouvoir calculer la MADD, le modèle $\mathcal{C}$ doit fournir pour chaque prédiction, soit une estimation de sa probabilité pour les modèles probabilistes (par exemple : réseaux bayésiens), soit un score de confiance pour les modèles non probabilistes (par exemple : arbres de décision), les deux étant représentés par une valeur comprise entre 0 et 1 inclus.

Par ailleurs, par simplification, nous utiliserons les termes *probabilités prédites* ou *probabilités* pour faire référence à la fois aux estimations de probabilité ou aux scores de confiance. Par exemple, avec un seuil de classification fixé à 0,5, un modèle prédit la réussite (valeur 1) s'il produit une *probabilité* supérieure à 0,5, sinon, il prédit l'échec.

3.2. EXPLICATIONS

Supposons que le modèle $\mathcal{C}$ prédise des probabilités de réussite comme illustrées en Figures 1a et 1b, pour deux groupes d'apprenants distincts G1 et G2. Ces deux groupes peuvent être des groupes distincts d'apprenants, c'est-à-dire qu'un apprenant ne peut pas appartenir aux deux groupes simultanément. Ces groupes peuvent être issus d'un même attribut (par exemple les hommes *vs.* les femmes pour l'attribut de genre) comme des groupes intersectionnels distincts (pour les attributs de genre et de handicap, les femmes avec un handicap forment un groupe distinct des hommes eux-mêmes avec un handicap).

Les histogrammes représentent alors la distribution des probabilités de réussite produites par le modèle $\mathcal{C}$, pour chacun de ces groupes. En Section 3.3, nous appellerons une telle distribution "vecteur de densité" des probabilités. Chaque barre verticale décrit la proportion d'apprenants ayant reçus la même probabilité de réussite. A titre d'exemple, sur ces histogrammes nous pouvons constater que les probabilités de G1 sont surtout situées entre 0 et 0,5, alors que celles de G2 sont plus élevées, entre 0,5 et 0,7 environ. Le modèle a donc tendance à donner de meilleures probabilités de réussite à G2 qu'à G1.

Ainsi, pour quantifier la différence de comportement du modèle $\mathcal{C}$ entre les deux groupes G1 et G2, la MADD a été conçue (Verger, Lallé *et al*., 2023) spécifiquement pour mesurer les différences entre les distributions que les histogrammes représentent. Nous en proposons une illustration en Figure 1c. Pour cela, nous avons illustré par des lignes courbes, les estimations continues des distributions représentées par les histogrammes à l'aide d'une méthode d'estimation de densité par noyau (ou *kernel density estimation*). La différence entre ces deux courbes, illustrée par une zone rouge dans laquelle le modèle ne produit pas les mêmes probabilités de réussite pour les deux groupes, est donc ce que la MADD cherche à quantifier. Nous présentons sa définition dans la section suivante. Il est important de souligner que la définition de la MADD ne s'appuie pas sur les estimations continues utilisées pour l'illustration mais bien sur les probabilités effectivement prédites par le modèle $\mathcal{C}$.

5. Traduisible en "Distance Absolue entre les Densités du Modèle".

3.3. DÉFINITION

Soient les vecteurs de densité $D^{G1} = [d_0^{G1}, d_1^{G1}, ..., d_m^{G1}]$ et $D^{G2} = [d_0^{G2}, d_1^{G2}, ..., d_m^{G2}]$, associés aux groupes G1 et G2 respectivement et illustrés par les Figures 1a et 1b. Les valeurs d_k avec $0 \leq k \leq m$ représentent les proportions de chaque probabilité discrète obtenue, c'est-à-dire la proportion de chaque barre verticale des Figures 1a et 1b, et où m correspond au nombre de probabilités discrètes considérées (Verger, Lallé *et al.*, 2023). Comme chaque vecteur représente la proportion totale des probabilités reçues dans chaque groupe, la somme de leurs éléments vaut toujours 1 respectivement. Ainsi, la MADD est définie comme suit :

$$\text{MADD}(D^{G1}, D^{G2}) = \sum_{k=0}^{m} |d_k^{G1} - d_k^{G2}| \tag{1}$$

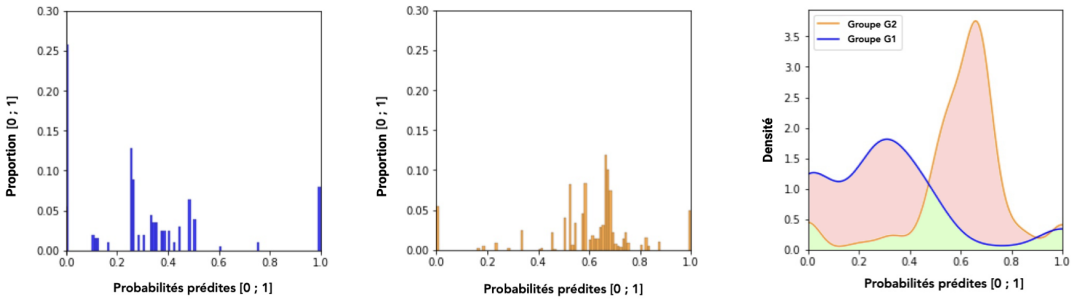
La MADD est par conséquent bornée entre 0 et 2. En effet, la MADD vaut 0 quand les deux vecteurs de densité sont identiques, c'est-à-dire que le modèle a le même comportement pour G1 et G2. A l'inverse, la MADD vaut 2 quand le modèle ne produit aucune probabilité commune entre les deux groupes. Une telle situation se produit par exemple quand le modèle donne une probabilité unique de p_i à tous les apprenants de G1 (i.e., proportion maximale pour une seule valeur de probabilité donnée) et une probabilité de p_j (avec $p_j \neq p_i$) à tous les apprenants de G2. Ainsi, pour n'importe quelles probabilités indexées par i et j :

$$\text{MADD}(D^{G1}, D^{G2}) = |d_i^{G1}| + |d_j^{G2}| = (1 + 1) = 2 \tag{2}$$

Par ailleurs, il existe un nombre optimal m de probabilités discrètes (ou de barres dans les histogrammes) à considérer pour le calcul de la MADD. Avec ce nombre optimal, la MADD présente des garanties théoriques de sa précision pour la mesure de la différence entre les deux distributions de chaque groupe, ce qui est montré dans Verger *et al.* (2024). Les expériences présentées dans les sections suivantes ont été réalisées avec ce nombre optimal. Cependant, la détermination de celui-ci n'est pas l'objet de cet article et nous renvoyons les lectrices et lecteurs au travail de Verger *et al.* (2024) quant à son choix et son usage.

4. ANALYSE DES GROUPES INTERSECTIONNELS

Dans cette section, nous revenons d'abord plus en détails sur des exemples de groupes intersectionnels en Section 4.1. Puis, nous présentons dans la Section 4.2 les différentes



(a) Groupe G1 (b) Groupe G2 (c) Approximation MADD

Figure 1 : Histogrammes des probabilités prédites pour deux groupes distincts (a, b) et représentation visuelle de la MADD (c)

approches possibles pour déterminer des groupes G1 et G2 intersectionnels (voir Equation 1) à analyser. Ces approches sont applicables avec toute mesure d'équité de groupe telle que la MADD et avec celles mentionnées en Section 2. Enfin, nous discutons du choix de ces approches en Section 4.3.

4.1. Exemples de groupes intersectionnels

Comme mentionné auparavant, un groupe intersectionnel est un groupe à l'intersection de plusieurs attributs. De tels groupes peuvent être représentés comme en Figure 2 (Yang *et al.*, 2020), où ils sont identifiés par les lettres A, B, C, D, E, F, G, H. À titre d'exemple, le groupe A inclut les femmes plus âgées avec un handicap. L'ensemble des femmes est alors l'union des groupes intersectionnels $A \cup B \cup E \cup F$, l'ensemble des personnes plus âgées $A \cup C \cup E \cup G$ et l'ensemble des personnes avec un handicap $A \cup B \cup C \cup D$.

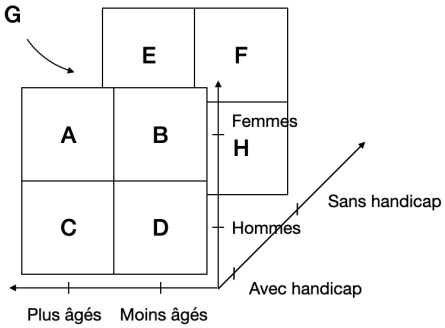


Figure 2 : Représentation des groupes intersectionnels avec trois attributs binaires (âge, genre et handicap)

4.2. Choix des groupes intersectionnels de comparaison

À l'inverse de l'analyse par attribut individuel, conduite usuellement dans la littérature (voir Section 2) et plus loin dans l'article (Section 6.1), consistant à comparer les groupes qui composent un attribut donné (par exemple, le groupe avec handicap et le groupe sans handicap pour l'attribut handicap), nous dénombrons deux manières de comparer les groupes intersectionnels, présentées ci-dessous.

4.2.1. Un groupe intersectionnel *vs.* le reste

Premièrement, il est possible de comparer un groupe intersectionnel (G1) avec le reste des apprenants dans les données (G2). Par exemple sur la Figure 2, le groupe A (femme plus âgées avec handicap) peut être comparé à l'union de tous les autres groupes (B à H). Ceci permet d'évaluer si un groupe donné est plus ou moins discriminé par rapport au reste des individus (hommes, femmes moins âgées sans handicap, femmes moins âgées avec handicap). Nous proposons une telle analyse en Section 6.2.4, conduisant à une visualisation sous forme d'histogrammes.

4.2.2. Un groupe intersectionnel *vs.* un autre groupe intersectionnel

Deuxièmement, il est possible de comparer un groupe intersectionnel (G1) vis-à-vis d'un autre groupe intersectionnel (G2). Par exemple, sur la Figure 2, nous pouvons comparer les groupes A et B entre eux. Cela permet d'évaluer si un groupe est plus ou moins discriminé

par rapport à un autre groupe spécifique. Nous avons également réalisé de telles analyses dans les Sections 6.2.2 et 6.2.3, conduisant à une visualisation sous forme de matrice.

4.3. DISCUSSION

Nous pouvons constater qu'avec la première approche, il est aisé de comparer tous les groupes intersectionnels un par un avec le reste des individus. De cette manière, nous pourrons ordonner ces résultats pour trouver les groupes intersectionnels conduisant aux résultats de MADD les plus sévères, et inversement. Cela permet donc d'identifier des groupes intersectionnels particuliers.

Par contre, avec la deuxième approche, il n'est pas forcément évident de savoir entre quels groupes intersectionnels effectuer la comparaison. Nous pouvons comparer chaque groupe intersectionnel avec tous les autres un par un (voir Figure 6). Mais certaines comparaisons, entre des groupes intersectionnels spécifiques, permettent de connaître l'impact discriminant d'un attribut sur un autre. En effet, en Figure 3, nous illustrons les deux types de comparaisons possibles entre deux attributs, en l'occurrence les attributs de genre et d'âge. Nous pouvons observer que dans la configuration de gauche en Figure 3, nous comparons l'équité algorithmique entre les groupes A-C, A-D, B-C, et B-D. De cette manière, les femmes sont toujours comparées aux hommes, en particulier les femmes avec une distinction d'âge et les hommes avec une distinction d'âge. Dans cette configuration, l'impact discriminant de l'âge sur le genre peut donc être mis en évidence. À l'inverse, dans la configuration de droite en Figure 3, les personnes plus âgées sont comparées avec celles moins âgées, en ajoutant une distinction de genre. Cette fois, c'est l'effet de l'appartenance d'un groupe du genre qui est évalué sur celui de l'âge. Ce sont donc les comparaisons A-B, A-D, C-B et C-D qui sont en jeu. Ainsi, seules les comparaisons A-D et B-C sont communes entre ces deux types de comparaisons et, par conséquent, selon que l'on souhaite évaluer dans un sens ou dans l'autre l'impact d'un attribut sur un autre, l'une ou l'autre des configurations sera préférée. Nous allons mettre en pratique cette approche en Section 6.2.3.

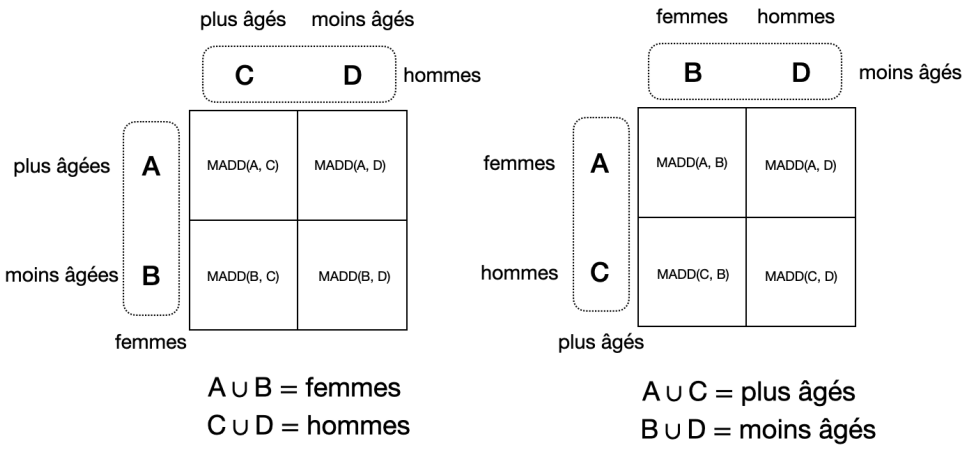


Figure 3 : Deux types de comparaisons pour évaluer l'impact discriminant d'un attribut sur un autre, avec l'exemple de deux attributs binaires (âge, genre)

5. EXPÉRIENCES

Dans les sections suivantes, nous allons présenter les données utilisées pour nos expériences (Section 5.1), les attributs sensibles considérés (Section 5.2) et les modèles étudiés pour les analyses (Section 5.3).

5.1. CORPUS DE DONNÉES OULAD

Nous conduisons nos analyses d'équité algorithmique à partir du jeu de données OULAD (*Open University Learning Analytics Dataset* - Kuzilek *et al*. (2017)). En effet, il s'agit d'un corpus anonymisé largement utilisé en éducation (Alhakbani et Alnassar, 2022), notamment pour la prédiction de la réussite à des cours en ligne. De plus, les données sont ouvertes, répondant spécifiquement à l'appel lancé à la communauté pour le développement de nouvelles approches sur des jeux de données ouverts (Hellas *et al*., 2018). Aussi, il contient des données de différents cours avec des profils d'apprenants variés, ce qui nous permet de répliquer nos expériences dans plusieurs contextes avec des populations différentes (autre appel lancé par Hellas *et al*. (2018)). Enfin, les données ont été collectées avec une attention particulière sur l'éthique et le respect de la vie privée.

Les cours du jeu de données OULAD ont été dispensés par *The Open University*, une université britannique à distance, qui propose des cours pouvant être suivis sans prérequis de manière indépendante ou dans le cadre d'un cursus universitaire. Les apprenants étaient inscrits entre 2013 et 2014 à au moins un des sept cours recensés dans le OULAD, dont trois en Sciences sociales et quatre en Science, Technologie, Ingénierie et Mathématiques (STIM). Ces informations sont regroupées dans le Tableau 1.

Tableau 1 : Informations sur le jeu de données OULAD

Cours	Domaine	Nombre d'apprenants
AAA	Sciences sociales	748
BBB	Sciences sociales	7 909
CCC	STIM	4 434
DDD	STIM	6 272
EEE	STIM	2 934
FFF	STIM	7 762
GGG	Sciences sociales	2 534

Le corpus contient des données démographiques et des données d'activité dans l'espace numérique de travail (ENT), avec initialement 32 593 échantillons (paire apprenant - cours). Nous avons utilisé les attributs présentés dans le Tableau 2 ainsi que la variable cible binaire « réussite/échec ». Seul l'attribut `nb_total_click` n'était pas immédiatement disponible dans le corpus et a été calculé par jointure et agrégation. Nous avons supprimé les échantillons avec des données manquantes et les valeurs de chaque attribut ont été normalisées entre 0 et 1 en prenant soin de ne pas appliquer de standardisation précisément pour garder les distributions de données originales pour l'analyse des discriminations algorithmiques.

5.2. ATTRIBUTS SENSIBLES ET SÉLECTION DES COURS

Nous ciblons dans nos expériences l'étude du caractère sensible des quatre attributs suivants : `genre`, `age`, `pauvrete` et `handicap`. Dans une recherche exhaustive d'attributs sensibles, il est tout à fait possible d'analyser avec notre méthode les discriminations algorithmiques relativement à tous les attributs disponibles dans un jeu de données, à condition

Tableau 2 : Attributs utilisés du jeu de données OULAD

Attribut	Type	Description
`genre`	binaire	genre de l'apprenant
`age`	ordinal	intervalle de l'âge de l'apprenant
`handicap`	binaire	indique si l'apprenant a déclaré un handicap
`dernier_diplome`	ordinal	dernier diplôme de l'apprenant à l'entrée du cours
`pauvrete`	ordinal	niveau de pauvreté du lieu d'habitation de l'apprenant
`nb_tentatives`	numérique	nombre de tentatives précédentes au cours
`credits`	numérique	nombre de crédits pour le cours étudié
`nb_total_click`	numérique	nombre total d'interactions de l'apprenant avec le cours

qu'ils soient binaires ou catégoriels pour la définition des groupes intersectionnels. Cependant, ici nous allons plutôt chercher à confirmer ou infirmer le caractère sensible de ces quatre attributs retenus. Par ailleurs, pour distinguer deux groupes G1 et G2 pour les attributs `pauvrete` et `age` dans leur analyse de manière individuelle (voir Section 6.1), nous utilisons un seuil de 50% de l'indice de pauvreté britannique (Kuzilek *et al.*, 2017) et nous distinguons, parmi les trois tranches d'âge disponibles dans les données ([0-35] : 13 815 individus, ou 69% ; [36-55] : 6 011 individus, ou 30% et [55+] : 138 individus, ou 18%), le groupe majoritaire ([0-35]) du groupe minoritaire (regroupement de [36-55] et [55+]).

Quant aux cours étudiés, parmi les sept disponibles, nous avons sélectionné dans le corpus le cours de Sciences sociales identifié "BBB" et le cours de STIM identifié "FFF". En effet, d'après une analyse des corrélations des attributs, notamment en Figure 4, ces deux cours ont présenté les plus fortes corrélations avec l'attribut `genre`, ce qui suggère une importance de celui-ci pour la prédiction de la réussite ou de l'échec par les modèles. De plus, ces deux cours ont aussi présenté de forts déséquilibres entre les deux groupes avec d'une part, une large majorité de femmes (91,2 %) dans le cours "BBB", et d'autre part, une majorité d'hommes dans le cours "FFF" (88,4 %). Par ailleurs, ces deux cours présentent l'avantage d'être dans deux domaines différents et d'avoir les effectifs d'apprenants les plus élevés (voir Tableau 1). Ainsi, ces deux cours sont de très bons candidats pour notre analyse des discriminations algorithmiques par rapport aux attendus de biais de genre notamment.

5.3. Modèles prédictifs de la réussite étudiés

Dans un souci de généralisation, nous avons sélectionné plusieurs types de modèles de classification pour nos analyses, respectivement à base de régressions, de distances, d'arbres et de probabilités : un modèle de régression logistique (LR), un modèle des k-plus proches voisins (KN), un arbre de décision (DT) et un classifieur naïf bayésien (NB). Le choix de ces modèles a été motivé par plusieurs raisons. Tout d'abord, les modèles susmentionnés sont largement utilisés dans le domaine de l'éducation (Alhakbani et Alnassar, 2022 ; Korkmaz et Correia, 2019), y compris avec le jeu de données OULAD (voir Section 2). D'autres modèles courants comme les machines à vecteurs de support n'ont pas été retenus car ils ne produisent pas d'estimations de probabilité (ou de scores de confiance) nécessaires pour effectuer notre analyse. Deuxièmement, bien que notre approche puisse être généralisée à d'autres modèles de complexité supérieure tels que des forêts aléatoires et des réseaux de neurones, nous avons privilégié les boîtes blanches et l'explicabilité sur l'optimisation que requiert ces modèles. Troisièmement, la prédiction de la réussite avec les données OULAD est un problème de prédiction à faible niveau d'abstraction, où l'utilisation de modèles prédictifs complexes conduirait à de moins bonnes performances et à un surapprentissage.

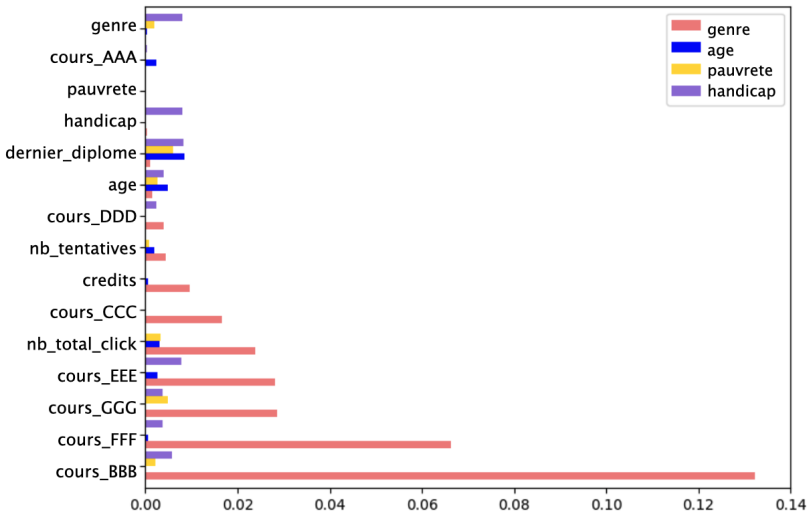


Figure 4 : Scores d'informations mutuelles (Kullback, 1959) entre les attributs de gauche avec ceux en légende

Concernant l'entraînement des modèles retenus, nous avons utilisé 70% des données pour le jeu d'entraînement et 30% pour le jeu de test, en gardant les mêmes proportions de réussite et d'échec dans les deux jeux de données à l'aide d'une sélection aléatoire stratifiée. Les modèles ont obtenu des précisions supérieures à la précision de référence (70% étant la proportion originale de réussite) allant jusqu'à 93%, à l'exception du classifieur NB (62%) qui, en revanche, a présenté des comportements intéressants pour l'analyse des discriminations algorithmiques. Nous soulignons que, contrairement aux études qui se concentrent sur l'apprentissage automatique, l'objectif ici n'est pas d'obtenir les meilleures performances prédictives mais d'illustrer l'intérêt de nos méthodes sur divers modèles. Ce point serait en revanche important dans le cas du déploiement réel d'un modèle prédictif. Enfin, les analyses et les évaluations avec le calcul de la MADD ont été réalisées sur le jeu de test. Ces évaluations peuvent être reproduites grâce aux données et au code documenté, disponibles à l'adresse indiquée à la note de pied de page n°3.

6. Résultats

Dans cette section, nous présentons les analyses d'équité algorithmique, d'une part, selon les attributs sensibles individuellement en Section 6.1, et d'autre part, selon les attributs sensibles simultanément grâce aux groupes intersectionnels en Section 6.2. Les premières ont été en partie reprises des travaux précédents (Verger, Bouchet *et al.*, 2023) afin d'être comparées aux deuxièmes.

6.1. Analyses d'équité par attributs sensibles individuels

Les résultats de MADD pour chaque modèle vis-à-vis de chaque attribut sensible sont présentés dans les Tableaux 3 et 4 pour le cours de Sciences sociales et de STIM respectivement. Ces tableaux se lisent comme suit : les meilleurs résultats de MADD par attribut (lecture en colonne) sont en **gras**, et les meilleurs résultats de MADD par modèle (lecture en ligne) portent une astérisque (*). Les valeurs les plus élevées représentant les discriminations les plus fortes par modèle sont quant à elles indiquées en rouge. Par ailleurs, les lectrices et

lecteurs peuvent consulter Verger, Bouchet *et al.* (2023) pour l'exploitation visuelle des résultats de la MADD pour obtenir davantage de détails.

6.1.1. Cours de Sciences sociales ("BBB")

Dans le Tableau 3, nous voyons par leur couleur rouge que les résultats de `pauvrete` sont les plus élevés pour trois modèles sur quatre (LR, KN et DT). Ces modèles produisent donc des prédictions de réussite qui diffèrent le plus en fonction de l'appartenance à l'un ou l'autre des groupes de `pauvrete`. L'`age` obtient des résultats similaires à `pauvrete`. Ce sont les deux attributs qui ont des moyennes de MADD les plus hautes (1,23 et 1,18).

En revanche, le `genre`, attendu comme l'attribut potentiellement le plus discriminant (voir Section 5.2), s'avère obtenir des valeurs inférieures à celles de `pauvrete` ou de `age`. Enfin, c'est le `handicap`, avec ces trois astérisques sur quatre, qui est l'attribut vis-à-vis duquel les trois modèles LR, KN et DT discriminent le moins. Il obtient d'ailleurs le moyenne la plus basse (0,82).

Tableau 3 : Résultats de la MADD pour le cours "BBB"

	Modèle	Attributs sensibles				Moyenne
		`genre`	`age`	`handicap`	`pauvrete`	
MADD	LR	1,72	1,80	1,57*	1,86	1,74
	KN	1,13	1,12	0,93*	1,13	1,08
	DT	**0,69**	**0,84**	**0,65***	**0,85**	0,76
	NB	**0,69***	1,14	1,13	0,87	0,96
	Moyenne	1,06	1,23	0,82	1,18	

6.1.2. Cours de STIM ("FFF")

Pour le cours de STIM, le Tableau 4 montre, qu'en revanche, `pauvrete` est l'attribut pour lequel les modèles discriminent le moins, avec la moyenne la plus basse (0,77). A l'inverse, l'attribut qui engendre le plus de discrimination pour trois modèles sur quatre est cette fois le `genre`, même si la moyenne pour l'`age` est également élevée.

Tableau 4 : Résultats de la MADD pour le cours "FFF"

	Modèle	Attributs sensibles				Moyenne
		`genre`	`age`	`handicap`	`pauvrete`	
MADD	LR	1,20	1,10	1,09	1.05*	1,11
	KN	1,05	0,96	0,79*	0,92	0,93
	DT	0,78	**0,68**	**0,60***	0,67	0,68
	NB	**0,53**	0,97	0,93	**0,44***	0,72
	Moyenne	0,89	0,93	0,85	0,77	

6.1.3. Bilan

Cette approche d'analyse d'équité algorithmique par attributs sensibles individuels nous a permis de déterminer leur pouvoir discriminant, soit en moyenne pour tous les modèles, soit pour chaque modèle en particulier. Cette approche a permis de mettre en exergue que `pauvrete` et `genre` étaient des attributs sensibles cruciaux dans le cours de Sciences

sociales ou de STIM, mais également que `age` avait toujours la moyenne de MADD la plus élevée dans les deux cours. Pour répondre à nos questions de recherche introduites en Section 1, nous allons maintenant étudier quels seraient les résultats d'équité si nous considérions plusieurs attributs sensibles ensemble.

6.2. Analyses d'équité par groupes intersectionnels

Nous nous intéressons ici aux discriminations intersectionnelles présentes dans les résultats, c'est-à-dire aux discriminations qui viennent de l'influence de plusieurs attributs sensibles simultanément. Pour rappel, un groupe intersectionnel est un groupe à l'intersection de plusieurs attributs. Ces analyses nous permettront de comparer l'influence d'attributs sensibles pris seuls dans la section précédente, avec l'influence d'attributs sensibles pris ensemble.

6.2.1. Cadre de ces analyses

Les analyses étant plus nombreuses du fait du nombre important de groupes intersectionnels possibles à partir de quatre attributs (i.e., `genre`, `age`, `handicap`, `pauvrete`), nous allons restreindre cette partie aux résultats du modèle DT (arbre de décision) et au cours "BBB". En effet, concernant le modèle DT, il est un bon candidat pour l'analyse plus approfondie des discriminations algorithmiques du fait de sa bonne précision et de ses bons résultats de MADD : il a obtenu la précision la plus élevée (93%) et il a obtenu les meilleures moyennes de MADD dans les deux cours de Sciences sociales et de STIM (voir les moyennes à droite dans les Tableaux 3 et 4). Par conséquent, il aurait été un modèle de choix dans un scénario réel, ce qui présente un intérêt pour les analyses.

De plus, parmi les résultats du modèle DT, nous nous concentrons sur le cours "BBB" car il contient les deux attributs ayant reçu la MADD la plus élevée parmi les deux cours, à savoir `age` (0,84) et `pauvrete` (0,85). Ainsi, par rapport à une valeur de référence de 0,85 qui est la valeur maximale obtenue par le DT, nous pourrons examiner si un groupe formé à partir, d'une part, de l'attribut `pauvrete` ou `age` et, d'autre part, d'un autre attribut, présente une discrimination encore plus élevée que `pauvrete` ou `age` seul, ou bien si cet autre attribut aura tendance à faire diminuer l'influence du premier.

Évaluer l'impact de `pauvrete` ou `age` sur un autre attribut sera l'objet des Sections 6.2.2 et 6.2.3. Puis, nous ferons une analyse de tous les groupes intersectionnels possibles en Section 6.2.4, pour une recherche exhaustive des discriminations algorithmiques.

6.2.2. L'influence de l'attribut `pauvrete` sur les autres attributs sensibles

Tout d'abord, nous examinons l'influence de `pauvrete` sur les autres attributs. En effet, comme il est individuellement celui qui a obtenu la valeur de MADD la plus élevée, et qu'il est donc l'attribut le plus discriminé ici, il est intéressant d'étudier si, évalué conjointement avec un autre attribut, il forme des groupes intersectionnels davantage discriminés. Les résultats sont présentés dans les Tableaux 5, 6 et 7.

Sur l'ensemble des trois tableaux, aucune valeur n'excède la valeur de référence 0,85, ce qui signifie qu'appartenir à un groupe d'un autre attribut n'aggrave pas les discriminations algorithmiques déjà observées pour l'attribut `pauvrete`, voire les réduit significativement puisque très proches de 0 : 0,11, 0,09, 0,06 ou encore 0,03. De plus, nous constatons que les résultats sur la diagonale (moins pauvre, moins pauvre)-(plus pauvre, plus pauvre) sont bien plus faibles que sur l'autre diagonale (plus pauvre, moins pauvre)-(moins pauvre, plus pauvre), ce qui est illustré avec des pointillés rouges ou oranges pour l'exemple dans le Tableau 5. Cela confirme le poids plus important, dans les discriminations algorithmiques, de

l'attribut `pauvrete` car la comparaison entre les groupes opposés de `pauvrete` cause une augmentation significative de la MADD alors même qu'elle est très faible pour les groupes opposés de `genre`, `handicap`, et `age`. Pour résumer, l'attribut `pauvrete` ne présente pas d'intersectionnalité avec les autres attributs, c'est-à-dire n'engendre pas de discriminations plus importantes par rapport à sa considération seule.

Tableau 5 : Résultats de la MADD pour les groupes intersectionnels de `genre` et `pauvrete`. En pointillés rouges est représentée la diagonale (moins pauvre, moins pauvre)-(plus pauvre, plus pauvre) et en pointillés orange la diagonale (plus pauvre, moins pauvre)-(moins pauvre, plus pauvre).

		Homme	
		Moins pauvre	Plus pauvre
Femme	Moins pauvre	0,11	0,46
	Plus pauvre	0,71	0,19

Tableau 6 : Résultats de la MADD pour les groupes intersectionnels de `handicap` et `pauvrete`

		Handicap	
		Moins pauvre	Plus pauvre
Non handicap	Moins pauvre	0,09	0,29
	Plus pauvre	0,72	0,37

Tableau 7 : Résultats de la MADD pour les groupes intersectionnels de `age` et `pauvrete`

		Moins âgé	
		Moins pauvre	Plus pauvre
Plus âgé	Moins pauvre	0,03	0,63
	Plus pauvre	0,69	0,06

6.2.3. L'influence de l'attribut `age` sur les autres attributs sensibles

L'autre attribut ayant obtenu la deuxième valeur la plus élevée de MADD, `age`, ne produit quant à lui pas les mêmes résultats. En effet, les Tableaux 8, 9 et 10 montrent que la MADD dépasse la valeur de référence (0,85) avec les trois autres attributs. Les groupes concernés sont le groupe des femmes plus âgées contre les hommes moins âgés (0,93), le groupe des personnes plus âgées déclarées sans handicap contre les personnes moins âgées déclarées avec handicap (0,91) et le groupe des personnes moins pauvres plus âgées contre les personnes plus pauvres moins âgées (1,15). Cela montre, par les mêmes croisements au niveau des diagonales, que pour l'attribut `pauvrete`, l'attribut `age` a un poids plus important que `genre`, `handicap` ou même `pauvrete` dans les discriminations algorithmiques intersectionnelles. Autrement dit, dans le cas analysé dans le Tableau 8, il y a une discrimination plus importante entre les femmes plus agées et les hommes plus âgés que les personnes (hommes et femmes) âgées *vs.* les personnes moins âgées. Pour résumer, l'attribut `age` présente une intersectionnalité avec les autres attributs, c'est-à-dire qu'il engendre des discriminations plus importantes par rapport à sa considération seule.

Tableau 8 : Résultats de la MADD pour les groupes intersectionnels de `genre` et `age`

		Homme	
		Moins âgé	Plus âgé
Femme	Moins âgé	0,06	0,66
	Plus âgé	0,93	0,32

Tableau 9 : Résultats de la MADD pour les groupes intersectionnels de `handicap` et `age`

		Handicap	
		Moins âgé	Plus âgé
Non handicap	Moins âgé	0,05	0,65
	Plus âgé	0,91	0,33

Tableau 10 : Résultats de la MADD pour les groupes intersectionnels de `pauvrete` et `age`

		Plus pauvre	
		Moins âgé	Plus âgé
Moins pauvre	Moins âgé	0,12	0,69
	Plus âgé	1,15	0,36

6.2.4. Tous les groupes intersectionnels

Enfin, nous examinons les discriminations algorithmiques avec tous les groupes intersectionnels possibles, à savoir seize (2^4) puisque nous avons quatre attributs sensibles binaires à disposition. Ces groupes intersectionnels sont recensés dans le Tableau 11, avec les alias que nous leur avons attribué pour l'analyse des figures à venir, ainsi que leur effectif. Nous pouvons soit calculer l'équité de chaque groupe contre le reste des apprenants (cf. Figure 5), soit calculer l'équité groupe contre groupe (cf. Figure 6), comme introduit en Section 4.

GROUPES INTERSECTIONNELS VS. LE RESTE. Dans la Figure 5, chaque barre verticale représente la MADD d'un groupe intersectionnel contre le reste des individus. La ligne pointillée rouge, indiquant la valeur de référence (0.85), montre qu'un groupe en particulier dépasse cette valeur : les hommes plus pauvres avec un handicap et plus âgés (alias A, combinaison 1-1-1-1). Du point de vue des effectifs, comme reporté dans le Tableau 11, le groupe A est très minoritaire puisque constitué de seulement 3 apprenants, mais obtient tout de même une MADD au moins deux fois supérieure à d'autres groupes minoritaires comparables comme les groupes B, E et F avec respectivement 10, 1 et 5 apprenants. La question des effectifs sera soulevée en discussion dans la Section 7. Par ailleurs, nous pouvons également observer quatre barres aux environs de 0,60 qui se distinguent des autres, les groupes B, E, I et M, dont le point commun est d'être des groupes dans lesquels les personnes ont déclaré un handicap. Ceci n'aurait pas pu être relevé sans cette analyse des discriminations intersectionnelles.

GROUPES INTERSECTIONNELS VS. GROUPES INTERSECTIONNELS. Dans la Figure 6, chaque case de la matrice contient un cercle dont le diamètre représente la valeur de MADD de la comparaison d'un groupe intersectionnel en ordonnées contre un groupe intersectionnel en abscisses. La lecture de la figure se fait donc de la gauche vers le bas. La diagonale étant les comparaisons de chaque groupe avec lui-même, les cases correspondantes contiennent bien une MADD de valeur nulle. De plus, en nommant à nouveau deux groupes

Tableau 11 : Alias de tous les groupes intersectionnels. Pour `genre`, les hommes sont désignés par la valeur 1 et les femmes par la valeur 0. Pour `pauvrete`, les plus pauvres sont désignés par la valeur 1 et les moins pauvres par la valeur 0. Pour `handicap`, ceux déclarés avec un handicap sont désignés par la valeur 1 et ceux qui n'ont déclaré aucun handicap par la valeur 0. Enfin, pour `age`, les personnes plus âgées sont désignées par la valeur 1 et les personnes moins âgées par la valeur 0.

Alias	`genre`	`pauvrete`	`handicap`	`age`	Effectif
A	1	1	1	1	3
B	1	1	1	0	10
C	1	1	0	1	44
D	1	1	0	0	44
E	1	0	1	1	1
F	1	0	1	0	5
G	1	0	0	1	35
H	1	0	0	0	44
I	0	1	1	1	31
J	0	1	1	0	48
K	0	1	0	1	234
L	0	1	0	0	514
M	0	0	1	1	17
N	0	0	1	0	27
O	0	0	0	1	210
P	0	0	0	0	323

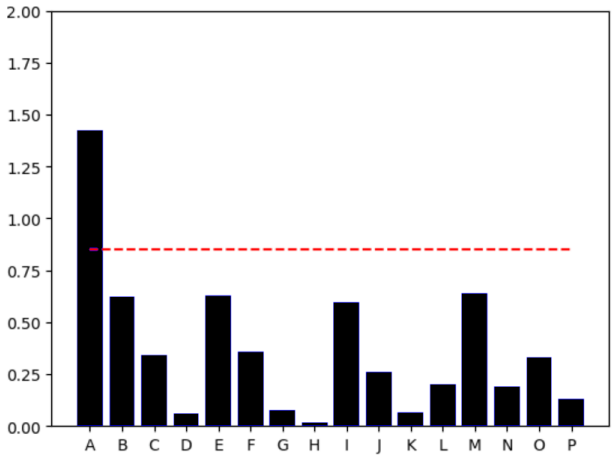


Figure 5 : Résultats de la MADD pour tous les groupes intersectionnels vis-à-vis du reste des apprenants

intersectionnels par G1 et G2, il est possible de comparer G1 avec G2 ou bien G2 avec G1. C'est pourquoi la matrice est symétrique. Nous avons également ajouté des couleurs pour préciser le sens des comparaisons : en lisant toujours de gauche à en bas et en prenant, par exemple, le groupe C en ordonnées, C présente un cercle vert avec A, signifiant qu'il est favorisé par le modèle par rapport à A, tandis qu'il présente un cercle rouge avec E, signifiant qu'il est plus discriminé par le modèle par rapport à E.

Dans cette matrice, nous pouvons constater deux choses. Premièrement, il y a 2 cases

avec une MADD maximale de 2 (i.e., cercle de diamètre maximal). Ce sont les comparaisons des groupes :

— (A) les hommes plus pauvres avec un handicap et plus âgés,

— (E) les hommes moins pauvres avec un handicap et plus âgés (`pauvrete` diffère),

et de :

— (A) les hommes plus pauvres avec un handicap et plus âgés,

— (M) les femmes moins pauvres avec un handicap et plus âgées (`genre` et `pauvrete` diffèrent).

Dans ces deux comparaisons A-E et A-M avec des effectifs respectifs de 3-1 et 3-17, nous tombons dans le cas où la MADD est maximale car le modèle ne produit aucune probabilité commune entre les deux groupes puisqu'il y a de toute façon très peu de probabilités concernées. A nouveau, la question des effectifs sera abordée dans la Section 7.

Deuxièmement, nous constatons que la ligne du groupe A semble celle ayant les valeurs globales les plus élevées, plus que le groupe E constitué pourtant d'un seul apprenant. Ce groupe A semble donc être le groupe envers lequel il est nécessaire d'être le plus vigilant dans le cours de Sciences sociales "BBB", du fait de son effectif très réduit et des probabilités très faibles qu'il reçoit (tous ses cercles sont rouges et de diamètre élevé). D'autres lignes avec une très grande majorité de cercles rouges peuvent également attirer notre attention comme les groupes B, C, I et J.

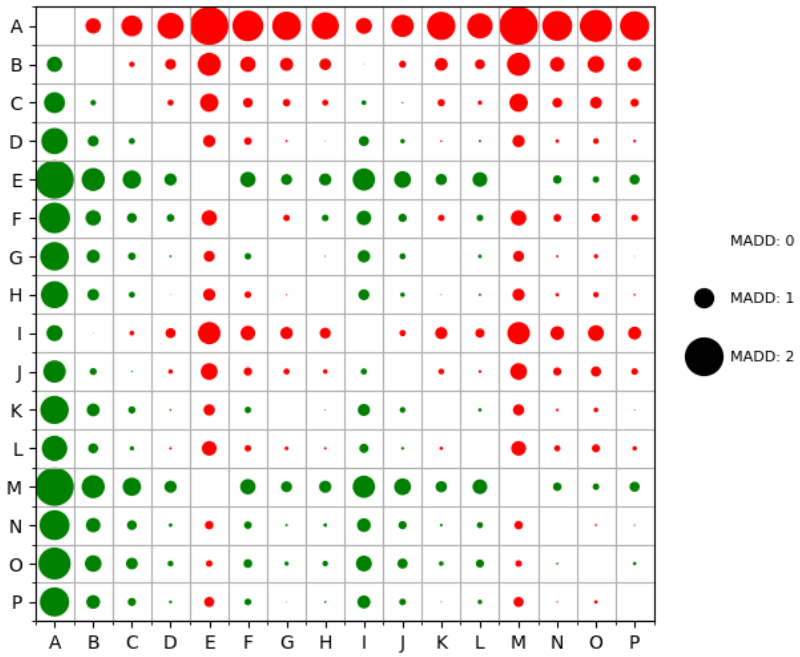


Figure 6 : Résultats de la MADD des groupes intersectionnels entre eux

6.2.5. Bilan

Grâce aux analyses d'équité algorithmique par groupes intersectionnels, nous avons pu comprendre de manière plus fine l'influence de chaque attribut sensible par leurs interactions avec les autres. Cela nous a également permis de découvrir des discriminations invisibles avec une analyse par attributs sensibles individuels, comme celles pour les groupes A, E et M, et en particulier celui concernant les hommes plus pauvres avec un handicap et plus âgés (groupe A). Une question peut donc émerger : comment ne pas passer à côté de toutes les

discriminations possibles ? Autrement dit, est-il possible de détecter automatiquement les bons groupes intersectionnels à analyser ? Ce point sera abordé ci-après.

7. DISCUSSION

Dans cette section, nous allons d'abord soulever les limites des données et modèles considérés (Section 7.1), puis les points clés des résultats obtenus avec les attributs sensibles individuels (Section 7.2) et avec les groupes intersectionnels (Section 7.3). Ensuite, nous discuterons du choix des attributs sensibles et des précautions à prendre vis-à-vis de leurs effectifs (Section 7.4). Enfin, nous proposerons quelques recommandations pratiques (Section 7.5).

7.1. DONNÉES ET MODÈLES ÉTUDIÉS

Le jeu de données OULAD contient peu d'attributs et presque tous catégoriels. Par conséquent, il présente une diversité et une variance très faibles dans les valeurs des attributs. Pour cette raison, tout modèle appliqué sur ces données aura des difficultés à distinguer avec précision les étudiants qui réussissent et ceux qui échouent aux cours. OULAD, qui est l'un des rares jeux de données ouverts comportant plusieurs attributs sensibles, présente donc un désavantage significatif sur la qualité limitée des données et la représentation réduite des individus, ce qui aurait pu permettre une évaluation plus robuste de l'équité algorithmique.

Aussi, dans nos expériences, nous nous sommes délibérément concentrés sur les résultats de la MADD pour les analyses d'équité. Cependant, pour des applications réelles, il est important de prêter attention à la fois à la performance et à l'équité afin de sélectionner des modèles pertinents pour leur usage. Par exemple, le modèle NB dans nos expériences pourrait être considéré comme équitable au regard de ses résultats de MADD, mais il présentait une précision très faible. En particulier, il avait des difficultés à prédire correctement la réussite ou l'échec des individus par rapport à tous les attributs, pas que sensibles, ce qui rendrait ce modèle inutile pour des cas réels, mais néanmoins intéressant ici pour notre analyse exploratoire. Notamment, nous n'avons effectué aucun ajustement des modèles (ou *fine tuning*) pour illustrer un cas général avec différents modèles. Il convient à l'inverse, d'étudier attentivement cette phase du processus. Nous recommandons donc d'utiliser la MADD pour des modèles qui montrent une performance prédictive satisfaisante concernant le problème en jeu, afin d'acquérir une compréhension plus fine de la façon dont ils se comportent et vis-à-vis de qui, et d'affiner leur sélection.

De plus, utiliser un modèle prédictif qui serait biaisé, pour une application réelle (et potentiellement à grande échelle), pourrait perpétuer ou produire des discriminations envers certains individus. L'enjeu est donc de systématiquement évaluer la performance et l'équité des modèles en profondeur pour être informé de ses potentiels biais, mais aussi pour anticiper les conséquences négatives de leur utilisation.

7.2. RÉSULTATS AVEC ATTRIBUTS SENSIBLES INDIVIDUELS

Les résultats au niveau des attributs sensibles ont permis de montrer qu'il n'y a pas de relation directe entre les biais dans les données étudiées en entrée et les biais dans les discriminations algorithmiques des modèles en sortie. En effet, malgré le biais de genre dans les données du cours de Sciences sociales, notre analyse montre que c'est un autre attribut sensible, pauvreté, qui est à l'origine des discriminations algorithmiques les plus importantes. En revanche, cela n'est pas le cas pour le cours de STIM qui, lui, présente des résultats de MADD plutôt alignés avec le biais de genre dans les données et des plus faibles discriminations algorithmiques selon la pauvreté.

7.3. Résultats avec groupes intersectionnels

Les résultats au niveau des groupes intersectionnels ont permis de mettre en lumière des discriminations qu'on ne pouvait voir avec l'analyse des attributs individuels seuls et de comprendre de manière plus fine l'influence de chaque attribut sur les discriminations grâce à leurs interactions avec les autres attributs. En effet, ces analyses ont permis de découvrir des groupes particuliers, dont celui des hommes plus pauvres avec un handicap et plus âgés, et de montrer que l'attribut d'âge engendrait une intersectionnalité avec les autres attributs, à l'inverse de l'attribut pauvreté. L'analyse par groupes intersectionnels est donc importante pour l'équité algorithmique, et a ainsi permis de répondre à la QR2, sur la découverte de discriminations algorithmiques supplémentaires en considérant les individus à l'intersection de plusieurs attributs sensibles.

7.4. Choix et effectifs des groupes intersectionnels

Dans la Section 4, nous avons présenté et discuté de deux approches pour étudier l'équité algorithmique vis-à-vis de plusieurs attributs sensibles. Ainsi, nous avons pu apporter une réponse à la QR1 qui était de savoir comment évaluer l'influence de plusieurs attributs sensibles simultanément. Nous pouvons en effet comparer un groupe intersectionnel avec le reste des individus, ou bien un groupe avec un autre groupe, dont certaines comparaisons de groupes peuvent faire ressortir l'influence d'un attribut sensible sur un autre.

Par ailleurs, il est à souligner que les groupes intersectionnels peuvent être formés à partir d'attributs binaires, comme dans cet article, ou catégoriels de manière générale, et que, pour des attributs numériques, une question de seuil se pose (comment faire des catégories pertinentes à partir de valeurs continues ?). Différents attributs conduisent à différents groupes intersectionnels. Il y a donc un choix à faire, qui reste empirique, des attributs à considérer, et donc des groupes intersectionnels à étudier.

En plus du choix des attributs sensibles, les effectifs des groupes intersectionnels représentent un point important de la fiabilité des mesures. En effet, à partir de l'importance de l'analyse par groupes intersectionnels, rappelée dans la section précédente, la question qui vient à se poser pour aller plus loin est : comment détecter toutes les discriminations possibles ? Autrement dit, comment déterminer le niveau d'intersectionnalité des groupes à considérer ? En effet, le niveau maximal d'intersectionnalité serait de considérer tous les attributs (sensibles ou non) d'un jeu de données s'ils sont catégoriels, ou de comparer les résultats des individus deux à deux s'il y a au moins un attribut numérique. Cependant, même si cela permet une recherche systématique des discriminations, en testant tous les attributs, il semble difficile de remonter au sens des discriminations quand les groupes intersectionnels sont trop fins, c'est-à-dire à l'intersection d'un trop grand nombre d'attributs, et par conséquent regroupant un trop faible nombre d'individus. Ce point sur les effectifs est en effet une perspective intéressante à examiner en lien avec l'équité : à partir de quelles proportions d'individus dans un jeu de données est-il pertinent de les comparer en termes d'équité algorithmique ?

7.5. Recommandations pratiques

Pour finir, nous proposons des recommandations pratiques pour utiliser la MADD à destination des chercheuses, chercheurs, développeuses et développeurs de modèles prédictifs en éducation. Tout d'abord, nous mettons à disposons une librairie Python, `maddlib` (lien à la note de pied de page n°4), pour partager des outils communs à l'analyse de l'équité. Ensuite, nous proposons les instructions suivantes, en 7 étapes, pour conduire une analyse d'équité, que ce soit avec des attributs individuels ou des groupes intersectionnels :

1. Choisir des modèles de classification binaire qui, en plus de leurs prédictions, produisent des estimations de probabilité ou des scores de confiance.
2. Transformer, si besoin, les attributs sensibles numériques en attributs binaires ou catégoriels.
3. Entraîner les modèles choisis.
4. En phase de test, séparer les probabilités prédites en fonction de chaque groupe à considérer, formé à partir d'un ou plusieurs attributs sensibles.
5. Calculer la MADD entre deux groupes d'intérêt (voir Section 6). Répéter le calcul autant de fois que de comparaison souhaitée.
6. Visualiser le comportement des modèles vis-à-vis des deux groupes d'intérêt et identifier graphiquement un traitement inégal ou un jugement stéréotypé des modèles (voir Verger, Bouchet *et al*. (2023) pour leurs explications).
7. Analyser les discriminations algorithmiques selon l'objectif recherché, dont voici des exemples principaux :
   — Identifier le modèle le plus équitable en moyenne ou par attribut, par une lecture en ligne des tableaux de résultats (voir Section 6).
   — Identifier l'attribut le plus équitable en moyenne ou par modèle, par une lecture en colonne des tableaux de résultats (voir Section 6).
   — Identifier le groupe le plus discriminé, par analyse visuelle (Verger, Bouchet *et al*., 2023), par un classement (voir Figure 5) ou une comparaison deux à deux (voir Figure 6).

Les étapes n°4 à 7 sont notamment facilitées par les outils de la librairie `maddlib`.

8. Conclusion

Dans cet article, nous avons conduit des analyses approfondies de l'équité algorithmique grâce à l'approche intersectionnelle. Nous avons mis en évidence les différences détectées selon les attributs sensibles individuels et les groupes intersectionnels présents dans le contexte éducatif du OULAD (Kuzilek *et al*., 2017). Ces analyses montrent l'importance d'inclure une approche intersectionnelle pour l'évaluation de l'équité de tout modèle prédictif, afin de détecter et de mieux comprendre certaines discriminations, au-delà des analyses traditionnellement menées sur les attributs sensibles seuls (voir Section 2).

Les deux principaux résultats de ces analyses sont les suivants. D'une part, les analyses ont permis de mettre en lumière des discriminations qu'on ne pouvait voir avec l'analyse des attributs individuels seuls. D'autre part, elles ont permis de comprendre de manière plus fine l'influence de chaque attribut sur les discriminations grâce à leurs interactions avec les autres attributs. Ces nouveaux résultats répondent à la QR2 posées en Section 1, grâce aux approches que nous avons développées pour la QR1, et ils renforcent la position de Verger, Bouchet *et al*. (2023) sur la nécessité d'analyser systématiquement les discriminations algorithmiques des modèles prédictifs pour chaque application éducative. Elles permettent en effet de fournir des éclairages sur les implications réelles de l'utilisation de ces modèles ainsi que sur les discriminations qui pourraient exister.

De plus, dans un soucis de reproductibilité des résultats et d'utilisation de notre méthode dans d'autres contextes éducatifs, nous mettons à disposition les données et le code documenté [6] ainsi que la librairie Python que nous avons nommé `maddlib` [7].

6. https://github.com/melinaverger/MADD
7. https://pypi.org/project/maddlib

Enfin, les perspectives pourront s'orienter vers l'étude de l'influence des effectifs des groupes intersectionnels sur les mesures d'équité, qui est un point crucial pour la fiabilité de ces mesures. Il est également intéressant d'explorer des méthodes d'atténuation des discriminations algorithmiques avec une approche intersectionnelle plutôt que de réduire ces discriminations par rapport à un seul attribut. Notamment, nous avons proposé une méthode de post-traitement utilisant la MADD (Verger, Fan *et al.*, 2023 ; Verger *et al.*, 2024) qui pourrait conduire à une expérience avec des groupes intersectionnels. Néanmoins, bien que toute méthode d'atténuation des discriminations algorithmiques (Caton et Haas, 2020 ; Mehrabi *et al.*, 2022) offre une première piste vers l'utilisation de systèmes plus équitables, elle ne prémunit pas contre une évaluation pas seulement informatique mais plus globale de l'équité des systèmes, avec la concertation des institutions, des enseignantes et enseignants, et des apprenantes et apprenants.

Références

Alhakbani, H. A., et Alnassar, F. M. (2022). Open learning analytics : a systematic review of benchmark studies using Open University learning analytics dataset (OULAD). *7th International Conference on Machine Learning Technologies (ICMLT)*, (p. 81-86). https://doi.org/10.1145/3529399.3529413

Baker, R. S., et Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *18*, 1052-1092. https://doi.org/10.1007/s40593-021-00285-9

Barocas, S., Hardt, M., et Narayanan, A. (2019). *Fairness and machine learning : limitations and opportunities* [http://www.fairmlbook.org]. MIT Press.

Buolamwini, J., et Gebru, T. (2018). Gender shades : intersectional accuracy disparities in commercial gender classification. *1st Conference on Fairness, Accountability and Transparency*, (p. 77-91). https://proceedings.mlr.press/v81/buolamwini18a.html

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., et Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(4209). https://doi.org/10.1038/s41598-022-07939-1

Caton, S., et Haas, C. (2020). Fairness in machine learning : a survey. *ACM Computing Surveys*, *56*(7), 1-38. https://api.semanticscholar.org/CorpusID:222208640

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex : a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, *140*, 139-167.

Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L., et Duy Le, T. (2022). How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics ? *British Journal of Educational Technology*, *53*(4), 822-843. https://doi.org/https://doi.org/10.1111/bjet.13217

Evans-Winters, V. E. (2021). Race and gender intersectionality and education. *Oxford Research Encyclopedia of Education*, *42*, 1-27.

Gardner, J., Brooks, C., et Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. *9th International Conference on Learning Analytics & Knowledge*, (p. 225-234). https://doi.org/10.1145/3303772.3303791

Gohar, U., et Cheng, L. (2023). A survey on intersectional fairness in machine learning : notions, mitigation, and challenges. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6619-6627. https://doi.org/10.24963/ijcai.2023/742

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., et Liao, S. N. (2018). Predicting academic performance : a systematic literature review. *23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, (p. 175-199). https://doi.org/10.1145/3293881.3295783

Hu, Q., et Rangwala, H. (2020). Towards fair educational data mining : a case study on detecting at-risk students. *13th International Conference on Educational Data Mining*, (p. 7).

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., et Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, (p. 656-666).

Kizilcec, R. F., et Lee, H. (2022). Algorithmic fairness in education. Dans *The ethics of artificial intelligence in education* (p. 174-202). Routledge.

Korkmaz, C., et Correia, A.-P. (2019). A review of research on machine learning in educational technology. *Educational Media International*, *56*(3), 250-267. https://doi.org/10.1080/09523987.2019.1669875

Kullback, S. (1959). *Information theory and statistics*. Wiley.

Kuzilek, J., Hlosta, M., et Zdrahal, Z. (2017). Open University learning analytics dataset. *Sci Data*, *4*(1), 1-8. https://doi.org/https://doi.org/10.1038/sdata.2017.171

Lee, H., et Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. *arXiv preprint arXiv :2007.00088*. http://arxiv.org/abs/2007.00088

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., et Galstyan, A. (2022). A survey on bias and fairness in machine learning. *arXiv :1908.09635 [cs]*. http://arxiv.org/abs/1908.09635

Pearl, J. (2009). *Causality* (2e éd.). Cambridge University Press.

Pessach, D., et Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys*, *55*(3), 1-44. https://doi.org/10.1145/3494672

Romero, C., et Ventura, S. (2020). Educational data mining and learning analytics : an updated survey. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, *10*(3), e1355. https://doi.org/10.1002/widm.1355

Verger, M. (2022). Investiguer la notion d'équité algorithmique dans les environnements informatiques pour l'apprentissage humain. *9ièmes RJC EIAH 2022 : Environnements Informatiques pour l'Apprentissage Humain*.

Verger, M., Bouchet, F., Lallé, S., et Luengo, V. (2023). Caractérisation et mesure des discriminations algorithmiques dans la prédiction de la réussite à des cours en ligne. *11ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain (EIAH 2023)*.

Verger, M., Fan, C., Lallé, S., Bouchet, F., et Luengo, V. (2023). A fair post-processing method based on the MADD metric for predictive student models. *st International Tutorial and Workshop on Responsible Knowledge Discovery in Education (RKDE 2023) at ECML PKDD 2023, Turino, Italy*. hal-04345451v1

Verger, M., Fan, C., Lallé, S., Bouchet, F., et Luengo, V. (2024). A comprehensive study on evaluating and mitigating algorithmic unfairness with the MADD metric. *Journal of Educational Data Mining (JEDM)*, *16*(1), 365-409.

Verger, M., Lallé, S., Bouchet, F., et Luengo, V. (2023). Is your model "MADD" ? A novel metric to evaluate algorithmic fairness for predictive student models. *16th International Conference on Educational Data Mining*.

Verma, S., et Rubin, J. S. (2018). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, (p. 1-7).

Yang, F., Cisse, M., et Koyejo, S. (2020). Fairness with overlapping groups. *Advances in Neural Information Processing Systems*, *33*.

Zambrano, A. F., Zhang, J., et Baker, R. S. (2024). Investigating algorithmic bias on bayesian knowledge tracing and carelessness detectors. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, (p. 349-359).