Caractérisation et mesure des discriminations algorithmiques dans la prédiction de la réussite à des cours en ligne

Mélina Verger, François Bouchet, Sébastien Lallé et Vanda Luengo

Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

Résumé. Les modèles prédictifs utilisés en EIAH peuvent produire des résultats biaisés et discriminants. Or, les mesures existantes pour les détecter sont seulement fondées sur l'égalité des performances prédictives entre différents groupes d'apprenant·e·s. Dans cet article, nous proposons une mesure objective des discriminations algorithmiques d'un modèle, ainsi qu'une méthode d'analyse visuelle pour caractériser ces discriminations. Nous démontrons l'application de notre méthode dans le cadre de la prédiction de la réussite à des cours en ligne, au moyen de données éducatives ouvertes. Nos résultats montrent la nécessité d'analyser systématiquement les discriminations algorithmiques issues des modèles afin de confirmer ou d'infirmer le caractère sensible de certains attributs.

Mots-clés : Équité algorithmique , métrique , attributs sensibles

Abstract. Preditive models used in intelligent learning environments can suffer from biased and unfair representation. However, existing fairness metrics that are meant to capture these issues are only based on the models' predictive performances. In this paper, we propose a novel fairness metric that measures to what extent the models behave unfairly. In addition, we provide a visualization-based analysis to qualify the types of unfair behaviors that are exhibited by the models. We apply our method on the success prediction task in online courses, with an open educational dataset. Our results highlight the need to systematically analyze unfair behaviors from the models in order to confirm or refute the sensitive nature of some attributes.

Keywords: Algorithmic fairness , metric , sensitive attributes

1 Introduction

La prédiction de la réussite à des cours en ligne a connu un intérêt important ces dix dernières années, comme le montre notamment une revue systématique relevant 357 articles publiés entre 2010 et mi-2018 sur le sujet [4]. Cependant, plusieurs exemples ont montré que les modèles prédictifs, au-delà des inégalités

et discriminations déjà présentes dans la réalité, pouvaient aussi être biaisés en faveur ou au détriment de certains groupes de la population, et ainsi amplifier voire produire de nouvelles discriminations [9,2,3,8]. Pour les distinguer des *discriminations historiques*, qui sont celles déjà présentes dans la société, nous nous réfèrerons à ces nouvelles discriminations issues de l'utilisation des modèles prédictifs sous le terme de *discriminations algorithmiques*. Ce constat alarmant sur l'utilisation des modèles prédictifs a conduit à une prise de conscience de leurs risques potentiels et notamment à la mise en place de directives par les instances de régulation [1].

Ainsi, il est devenu nécessaire d'analyser les discriminations apprises et produites par les modèles prédictifs. En particulier, une telle analyse doit permettre d'identifier envers quels groupes spécifiquement les modèles produisent les résultats les plus biaisés, pour pouvoir appréhender les implications réelles de l'utilisation de ces modèles. Dans la littérature (voir section 2), l'approche systématiquement employée pour quantifier les discriminations algorithmiques d'un modèle consiste à identifier au préalable les groupes de personnes potentiellement concernés, puis de comparer les performances prédictives (e.g. précision, F1-score) du modèle en fonction de ces groupes distincts. Un modèle est alors considéré comme ayant un comportement discriminant envers un groupe s'il ne produit pas des performances prédicitives similaires à celles obtenues pour les autres groupes.

Pour autant, évaluer si un modèle produit des performances prédictives similaires entre différents groupes (i.e. le même nombre d'erreurs) ne tient pas compte du fait qu'il puisse produire des erreurs plus sévères (et dont les implications seraient plus néfastes) pour un groupe que pour un autre, en dépit d'un nombre d'erreurs identique. C'est pourquoi nous proposons une nouvelle méthode d'analyse des discriminations algorithmiques visant à quantifier la sévérité de ces erreurs grâce à une nouvelle métrique nommée *Model Absolute Density Distance* (MADD) et fondée sur la caractérisation de ces discriminations via une analyse visuelle dérivée de cette métrique. Cette nouvelle métrique et l'analyse visuelle associée sont indépendantes des performances prédictives des modèles, pour permettre de quantifier les discriminations algorithmiques uniquement. Cette approche est particulièrement destinée aux chercheurs·euses et développeurs·euses de modèles prédictifs en éducation.

Nous appliquons notre approche sur un cas d'usage de prédiction de la réussite à des cours en ligne avec des données éducatives ouvertes, dans un souci de reproductibilité des expériences, et avec quatre types de modèles de classification binaire très courants pour cette tâche [1], dans un souci de généralisation. Nous mettons à disposition les données et le code documenté, permettant la réplication et l'utilisation de notre méthode dans d'autres contextes, à l'adresse suivante : `https://github.com/melinaverger/MADD`.

1. 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*.

2 État de l'art

D'après différentes revues de littérature [4,6], la prédiction de la réussite à des cours en ligne est le plus souvent représentée par un problème de classification binaire (e.g. réussite/échec). Plusieurs types de modèles sont couramment utilisés, tels que les réseaux bayésiens, les arbres de décisions, ou la méthode des k plus proches voisins. En expérimentant notre approche avec quatre types de modèles de classification binaire très courants en éducation [4,6] et en particulier sur le corpus OULAD [7,1], notre contribution s'inscrit dans la continuité de ces pratiques, et nous détaillerons ces choix dans la section 4. Le corpus OULAD a notamment été utilisé dans plusieurs travaux de prédiction connexes (i.e. réussite/échec, abandon/complétion des cours) [1], mais sans analyse des discriminations algorithmiques dans les prédictions.

Par ailleurs, quelques travaux en éducation [3,5,8], menés aux Etats-Unis, ont cherché à évaluer les discriminations apprises par des modèles de classification binaire. Or, comme abordé en introduction, au lieu de déterminer les groupes d'apprenant·e·s subissant des discriminations algorithmiques à partir des prédictions des modèles, les auteurs choisissent d'abord les groupes d'apprenant·e·s qu'ils jugent à risque, puis comparent les performances des modèles entre ces différents groupes. [3] les a par exemple comparé par rapport au genre, et [5] et [8] par rapport au genre et à l'origine ethnique des apprenant·e·s [2]. Le choix préalable de ces caractéristiques à risque, aussi appelées *attributs sensibles*, est souvent basé sur les résultats d'études de sciences sociales ou sur les caractéristiques mises en avant par des lois anti-discrimination (e.g. genre, origine ethnique, religion, handicap). En revanche, cette approche ne peut pas rendre visible les potentielles discriminations algorithmiques par rapport à d'autres attributs sensibles non sélectionnés par les auteurs. En effet, les discriminations apprises par un modèle peuvent être différentes de celles présupposées, puisqu'elles dépendent non seulement de la nature et de la représentation des données utilisées, mais aussi de la manière dont le modèle apprend de celles-ci. C'est pourquoi nous proposons une nouvelle approche fondée à la fois sur la quantification et la caractérisation des discriminations algorithmiques (voir section 3), afin de confirmer ou d'infirmer *a posteriori* le caractère *sensible* de certains attributs.

Enfin, notre approche repose exclusivement sur l'analyse des discriminations algorithmiques des modèles indépendamment de leurs performances prédictives, et diffère en cela des travaux existants sur l'équité des modèles prédictifs en éducation [3,5,8]. En effet, ces travaux comparent par exemple la différence de précision des modèles entre les différents groupes, ou le taux de bonnes prédictions par rapport au taux de mauvaises prédictions. Cependant, évaluer si ces performances prédictives sont égales à travers les groupes n'est pas synonyme d'absence de discrimination : un modèle peut produire des erreurs en même quantité, mais qui peuvent être très nuisibles pour un groupe et très peu pour

2. 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.

l'autre. Dans la mesure où différentes formes de discrimination algorithmique existent [12,10], nous avons développé une mesure indépendante de la performance prédictive, la MADD, capable de les quantifier.

3 Méthode d'analyse des discriminations algorithmiques

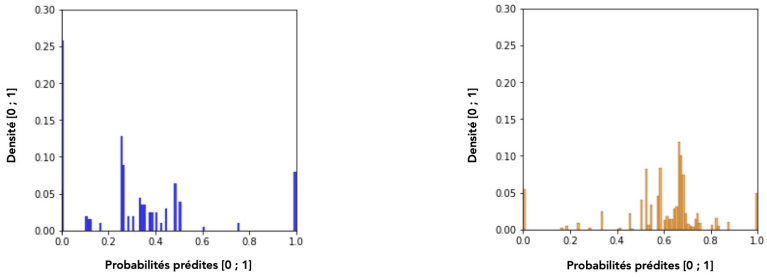
Dans cette section, nous présentons notre méthode d'analyse des discriminations algorithmiques des modèles prédictifs de classification binaire, tout d'abord via l'explication de l'analyse visuelle qui la compose (partie 3.1) puis via la définition de la nouvelle métrique MADD (partie 3.2).

Avant cela, considérons des modèles de classification binaire pour la prédiction la réussite à un cours en ligne. Pour appliquer notre méthode, chaque modèle doit fournir pour chaque prédiction (i.e. 0 pour échec ou 1 pour réussite) soit une estimation de sa probabilité pour les modèles probabilistes (e.g. réseaux bayésiens) soit un score de confiance pour les modèles non probabilistes (e.g. arbres de décision), les deux étant représentés par une valeur comprise entre 0 et 1. Par simplification, nous utiliserons les termes *probabilités prédites* ou *probabilités* pour faire référence à la fois aux estimations de probabilité et aux scores de confiance. Par exemple, avec un seuil de classification fixé à 0,5, un modèle prédit la valeur 1 (réussite) s'il produit une *probabilité* supérieure à 0,5, et prédit 0 (échec) sinon.

3.1 Analyse visuelle des discriminations algorithmiques

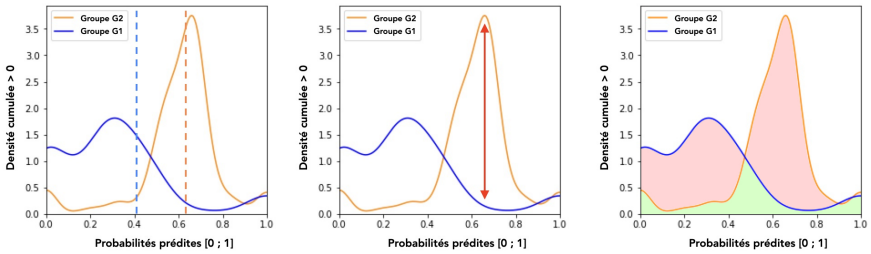
Au lieu de nous intéresser seulement aux prédictions 0 ou 1 que produisent les modèles, comme comparées dans les travaux cités en section 2, nous étudions de manière plus fine leurs probabilités prédites. Pour cela, nous étudions les fréquences avec lesquelles les modèles attribuent ces probabilités, en particulier celles associées à la prédiction 1 (réussite). Par exemple, dans la Figure 1, les histogrammes montrent pour un modèle donné la distribution des probabilités liées à la réussite pour deux groupes d'apprenant·e·s distincts, G1 et G2 (e.g. les apprenant·e·s déclarés avec un handicap (G1) et les apprenant·e·s déclarés sans handicap (G2)). Chaque barre verticale sur les Figures 1a et 1b représente la proportion d'apprenant·e·s ayant reçue la même probabilité de réussite. Nous appellerons par la suite une telle distribution *vecteur de densité* des probabilités.

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, principalement 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 faciliter l'analyse visuelle de ces histogrammes, difficilement interprétables en raison des nombreuses variations observées, nous proposons d'appliquer un lissage par une méthode d'estimation de densité par noyau (ou *kernel density estimation*). Nous utilisons plusieurs noyaux gaussiens pour approximer les distributions discrètes montrées dans les histogrammes et le coefficient de lissage est calculé automati-



(a) Groupe G1 (b) Groupe G2

FIG. 1 – Histogrammes des probabilités prédites pour deux groupes distincts.



(a) Traitement inégal (b) Jugement stéréotypé (c) MADD approximée

FIG. 2 – Représentations visuelles des formes de discriminations algorithmiques (a, b) et de la MADD (c). Les lignes en pointillées représentent les moyennes des distributions de probabilités.

quement par la règle de Scott [3] qui prend en compte le nombre d'échantillons et le nombre d'attributs présents dans le jeu de données. La Figure 2 présente le résultat d'un tel lissage pour les histogrammes de la Figure 1. Le passage de la distribution discrète à la distribution lissée change la densité en densité cumulée (axe des ordonnées), ce qui explique pourquoi les valeurs observées peuvent être supérieures à 1 dans la Figure 2.

Ce lissage permet de caractériser deux formes de discriminations algorithmiques :

1. Le *traitement inégal* : un modèle peut donner en moyenne de meilleures probabilités à un groupe qu'à un autre (Figure 2a), ce qui traduit le favoritisme du modèle pour un groupe par rapport à l'autre.
2. Le *jugement stéréotypé* : un modèle peut donner à de nombreux apprenant·e·s dans un groupe la même probabilité, ce qui traduit un comportement répétitif et invariant, caractéristique du stéréotype (Figure 2b).

Nous cherchons donc, dans la partie 3.2 suivante, à quantifier ce qui correspond à la zone rouge en Figure 2c, zone où un modèle ne produit pas les mêmes

3. Voir la documentation Python de `scipy.stats.gaussian_kde`.

probabilités de réussite pour les deux groupes. Il est cependant important de noter que cette quantification n'utilise pas les résultats du lissage effectué, qui ne vise qu'à faciliter l'analyse visuelle, mais bien les probabilités qui sont effectivement prédites par les modèles.

3.2 Métrique *Model Absolute Density Distance* (MADD)

Nous définissons la métrique MADD [4] comme suit. Posons 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, où m correspond au nombre de valeurs discrètes possibles que les probabilités de réussite peuvent prendre [11]. Comme chaque vecteur représente la fréquence des probabilités, la somme de ses éléments vaut toujours 1. Ainsi :

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

La MADD est 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 apprenant·e·s de G1 (i.e. densité maximale pour une seule valeur de probabilité donnée) et une probabilité de p_j (avec $p_j \neq p_i$) à tous les apprenant·e·s de G2. Ainsi, pour n'importe quelles probabilités données, indexées par i et j :

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

4 Expériences

4.1 Corpus de données OULAD

Nous expérimentons notre méthode sur le jeu de données OULAD (*Open University Learning Analytics Dataset*) [7]. Il s'agit en effet d'un corpus anonymisé largement utilisé en éducation [1], y compris pour la prédiction de la réussite à des cours en ligne ; 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 [4] ; et il contient des données de différents cours avec des profils d'apprenant·e·s variés, ce qui nous permet de répliquer nos expériences dans plusieurs contextes avec des populations différentes (autre appel lancé par [4]). De plus, 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

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

Table 1 – Attributs utilisés du jeu de données OULAD.

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

suivis sans prérequis de manière indépendante ou dans le cadre d'un cursus universitaire. Les apprenant·e·s étaient inscrit·e·s 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).

Le corpus contient des données démographiques et des données d'activité dans l'espace numérique de travail (ENT), avec 28 785 échantillons (paire apprenant·e - cours). Nous avons utilisé les attributs présentés en Table 1 ainsi que la variable cible binaire "Réussite"/"Echec". 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.

4.2 Attributs sensibles d'étude et sélection des cours

Pour nos expériences, nous ciblons l'étude du caractère sensible aux 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. Par exemple, les attributs `dernier_diplome`, `nb_tentatives`, `credits` et `nb_total_click` pourraient être pris en compte de manière complémentaire pour évaluer, en plus d'informer respectivement sur l'état des connaissances préalables, l'expérience du cours, son attractivité et l'intensité de l'activité des apprenant·e·s en son sein qui ne sont pas des informations à risque pour la prédiction du succès, leur sensibilité sans être démographiques. Nous allons ici plutôt confirmer ou infirmer le caractère sensible des quatre attributs retenus. Par ailleurs, pour distinguer deux groupes G1 et G2 pour `pauvrete` et `age` respectivement, nous utilisons un seuil de 50% de l'indice de pauvreté britannique (voir [7]) et nous distinguons, parmi les trois tranches d'âge disponibles dans les données ([0-35], [35-55] et [55+]), le groupe majoritaire ([0-35]) du groupe minoritaire (regroupement de [35-55] et [55+]).

Quant aux cours étudiés, nous avons sélectionné un cours en sciences sociales, identifié "BBB" dans le corpus, et un cours en STIM, identifié "FFF". En effet,

d'après une analyse des corrélations des attributs, ces deux cours ont présenté les plus fortes corrélations avec l'attribut `genre`, ce qui suggère une importance de cet attribut 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 le constituant dans le corpus, c'est-à-dire une large majorité de femmes (91.2 %) dans le cours "BBB", et à l'inverse une majorité d'hommes dans le cours "FFF" (88.4 %). Le choix de ces deux cours était donc pertinent pour notre analyse des discriminations algorithmiques par rapport aux attendus de biais de genre.

4.3 Modèles prédictifs de la réussite

Dans un souci de généralisation, nous expérimentons notre approche avec plusieurs types de modèles de classification, respectivement à base de régression, 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 [6,1] et 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 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.

Nous avons entraîné les modèles en utilisant 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. 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 à nouveau que contrairement aux études d'apprentissage automatique classiques, l'objectif ici n'est pas d'obtenir les meilleures performances prédictives mais de présenter l'intérêt de notre méthode sur divers modèles. Puis, nous avons calculé la MADD et réalisé les analyses visuelles sur le jeu de test.

5 Résultats

5.1 Cours de sciences sociales ("BBB")

Le Tableau 2 présente les résultats de la MADD pour chaque modèle et chaque attribut sensible dans le cours de sciences sociales. 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 en rouge. Le Tableau 2 montre ainsi que `handicap` est l'attribut vis-à-vis duquel trois modèles sur quatre (LR, KN et DT) discriminent le moins, avec la moyenne la plus basse à 0.82. À l'inverse, pour identifier les apprenant·e·s les plus discriminé·e·s, nous nous intéressons aux attributs qui induisent une valeur de MADD la plus élevée à travers tous les modèles. L'attribut `pauvrete` s'avère

TABLE 2 – Résultats de la MADD pour le cours "BBB".

	Modèle	Attributs sensibles				
		`genre`	`age`	`handicap`	`pauvrete`	Moyenne
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	

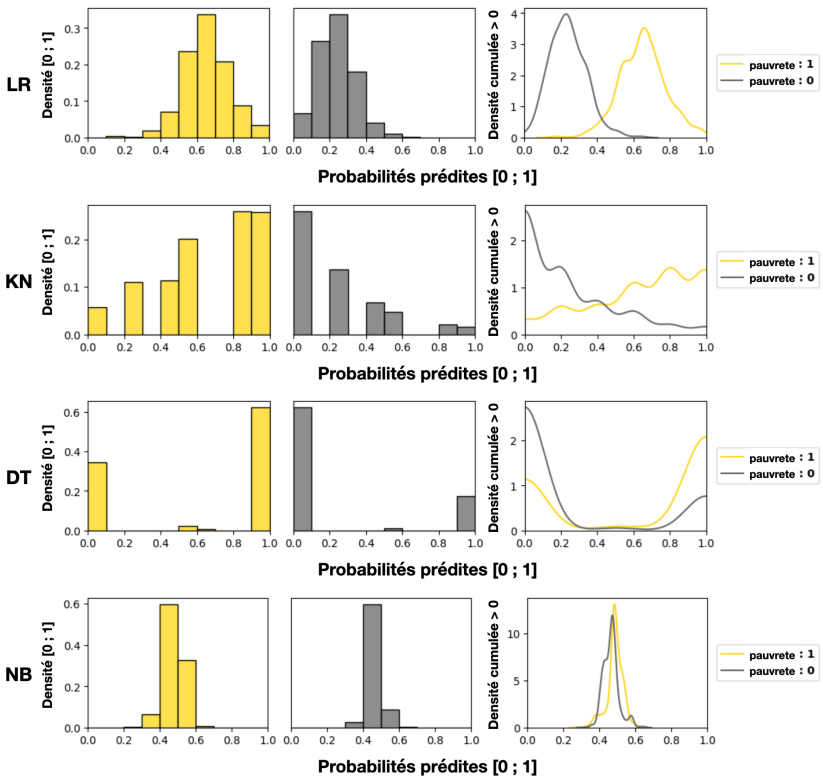


FIG. 3 – Analyse visuelle des vecteurs de densité pour le cours "BBB". Les barres des histogrammes ont été épaissies à des fins de visualisation.

être le plus discriminant pour trois modèles sur quatre (LR, KN et DT). Ainsi, pour caractériser les formes de discriminations algorithmiques vis-à-vis de cet attribut et identifier le groupe de `pauvretre` spécifiquement le plus discriminé, nous observons en Figure 3 que LR, KN et DT ont appris un "traitement inégal" (i.e. écart de moyennes significatif) en défaveur du groupe 0, les personnes vivant dans des régions les moins pauvres, obtenant les moins bonnes probabilités de réussite. Par conséquent, dans ce cours, les apprenant·e·s vivant dans des régions plus aisées sont les plus négativement discriminé·e·s (traitement inégal) par

TABLE 3 – Résultats de la MADD pour le cours "FFF".

	Modèle	Attributs sensibles				
		`genre`	`age`	`handicap`	`pauvrete`	Moyenne
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	

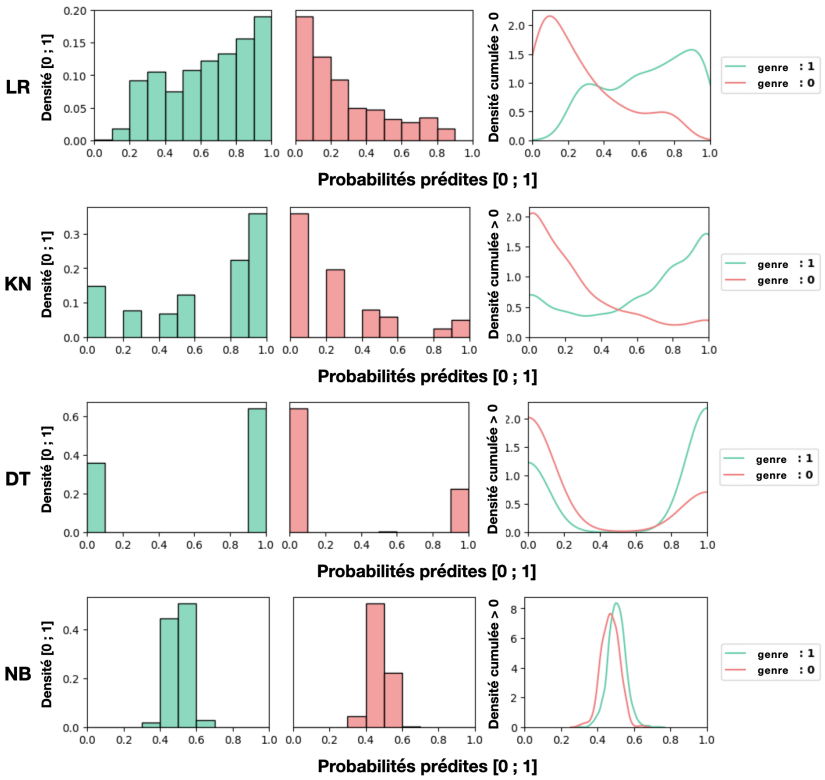


FIG. 4 – Analyse visuelle des vecteurs de densité pour le cours "FFF". Les barres des histogrammes ont été épaissies à des fins de visualisation.

la majorité des modèles, ou inversement ces modèles discriminent positivement les apprenant·e·s des régions défavorisées. Le NB a lui appris un comportement plus équilibré tandis que le DT et le KN montrent un comportement très stéréotypé avec seulement quelques pics distincts (voir histogrammes en Figure 3, qui montrent par exemple que pour le groupe 1, le KN donne 0.6, 0.8 ou 1.0 à 70 % des apprenant·e·s). Cela reflète bien le fonctionnement inhérent à ces deux modèles.

5.2 Cours de STIM (FFF)

Pour le cours de STIM, le Tableau 3 montre qu'en revanche `pauvrete`, avec la moyenne la plus basse (0.77), est l'attribut vis-à-vis duquel les modèles discriminent le moins. 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. Ainsi, d'après la Figure 4, les comportements des modèles révèlent majoritairement une inégalité de traitement en défaveur du groupe 0, les femmes. Par conséquent, dans ce cours, les femmes représentent le groupe le plus négativement discriminé (traitement inégal) par les modèles.

6 Conclusion

Les résultats conduisent à deux conclusions principales. Premièrement, 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. Ainsi, 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 a l'origine des discriminations algorithmiques les plus importantes. Deuxièmement, les analyses visuelles permettent de montrer que chaque modèle, même entraîné sur des données identiques, produit des discriminations algorithmiques différentes, observables à travers la variabilité des distributions dans les Figures 3 et 4.

Ainsi, ces conclusions démontrent la nécessité d'analyser systématiquement les discriminations algorithmiques des modèles prédictifs pour confirmer ou infirmer le caractère *sensible* de certains attributs. Cependant, il faut souligner qu'en pratique la MADD doit être utilisée avec des modèles présentant des performances prédictives satisfaisantes pour être utilisés dans des applications réelles. Dans notre cas, ceci exclurait le modèle NB qui, bien que présentant de bons résultats de MADD pour les attributs sensibles, en présentait aussi pour tous les attributs du corpus, le rendant par conséquent mauvais prédicteur de la réussite ou de l'échec (d'où sa plus faible précision en partie 4.3).

Notre méthode peut être utilisée de la même manière que dans cet article pour (1) quantifier les discriminations algorithmiques selon différents attributs, (2) caractériser la nature de ces discriminations, et (3) identifier spécifiquement les groupes les plus discriminés par les modèles, dans différents contextes. Notre travail vise plus largement à encourager la communauté à analyser les comportements des modèles actuellement intégrés dans des EIAH, et

nous mettons à disposition les données et le code documenté à l'adresse suivante : `https://github.com/melinaverger/MADD`. Nos prochains travaux visent à la généralisation de la définition de la MADD pour prendre en compte l'influence de plusieurs attributs simultanément au lieu d'un seul.

Références

1. Alhakbani, H.A., Alnassar, F.M. : Open Learning Analytics : A Systematic Review of Benchmark Studies using Open University Learning Analytics Dataset (OULAD). In : 7th International Conference on Machine Learning Technologies (ICMLT). pp. 81–86. ACM, New York, NY, USA (2022)
2. Buolamwini, J., Gebru, T. : Gender Shades : Intersectional Accuracy Disparities in Commercial Gender Classification. In : 1st Conference on Fairness, Accountability and Transparency. pp. 77–91. PMLR (2018)
3. Gardner, J., Brooks, C., Baker, R. : Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In : 9th International Conference on Learning Analytics & Knowledge. pp. 225–234. ACM, Tempe AZ USA (Mar 2019)
4. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V.V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., Liao, S.N. : Predicting academic performance : A systematic literature review. In : 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education. p. 175–199. ACM, New York, NY, USA (2018)
5. Hu, Q., Rangwala, H. : Towards Fair Educational Data Mining : A Case Study on Detecting At-risk Students. In : 13th International Conference on Educational Data Mining. p. 7 (2020)
6. Korkmaz, C., Correia, A.P. : A review of research on machine learning in educational technology. Educational Media International **56**(3), 250–267 (2019)
7. Kuzilek, J., Hlosta, M., Zdrahal, Z. : Open university learning analytics dataset. Sci Data 4 **170171** (2017)
8. Lee, H., Kizilcec, R.F. : Evaluation of Fairness Trade-offs in Predicting Student Success. arXiv :2007.00088 [cs] (Jun 2020)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. : A Survey on Bias and Fairness in Machine Learning. arXiv :1908.09635 [cs] (Jan 2022)
10. Verger, M. : Investiguer la notion d'équité algorithmique dans les environnements informatiques pour l'apprentissage humain. In : 9ièmes RJC EIAH 2022 : Environnements Informatiques pour l'Apprentissage Humain (2022)
11. Verger, M., Lallé, S., Bouchet, F., Luengo, V. : Is Your Model "MADD" ? A Novel Metric to Evaluate Algorithmic Fairness for Predictive Student Models. In : 16th International Conference on Educational Data Mining (2023)
12. Verma, S., Rubin, J.S. : Fairness definitions explained. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) pp. 1–7 (2018)