Métriques d’équité en Apprentissage Automatique et droit de l’Union Européenne en matière de non-discrimination

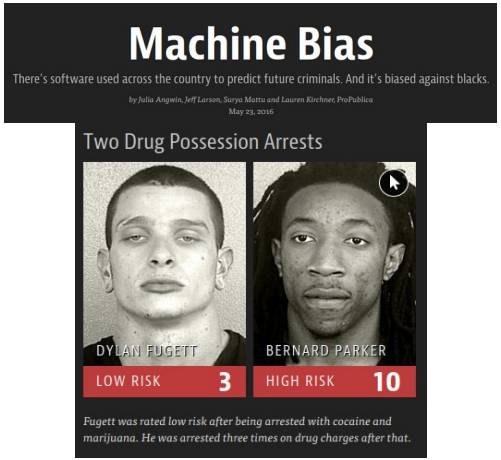
Magali Legast, Yasaman Yousefi, Lisa Koutsoviti Koumeri,
Axel Legay, Christoph Schommer, Koen Vanhoof

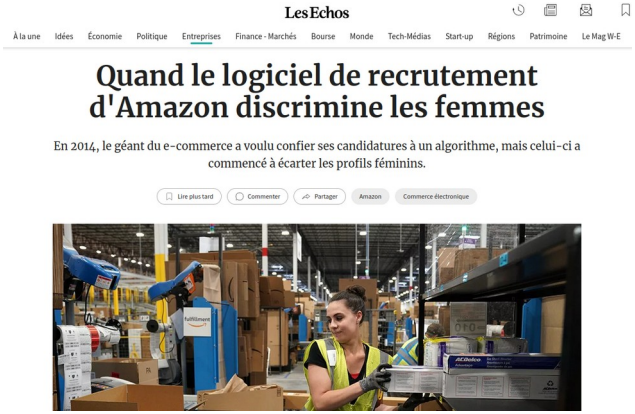
Conférence Nationale en Intelligence Artificielle
Strasbourg, France
3 - 5 juillet 2023

UCLouvain
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
UHASSELT KNOWLEDGE IN ACTION
uni.lu UNIVERSITÉ DU LUXEMBOURG

L’équité, nécessaire, complexe et interdisciplinaire

- Décisions algorithmiques empreintes de biais et résultats discriminatoires
- Comment allier réalité algorithmique et contraintes légales ?





Pays-Bas. Scandale des allocations familiales : un avertissement qui montre l’urgence d’interdire les algorithmes racistes

Le gouvernement néerlandais risque d’aggraver la discrimination raciale en continuant à utiliser des algorithmes non réglementés dans le secteur public, a déclaré Amnesty International dans un nouveau rapport accablant, où elle analyse l’affaire scandaleuse des allocations familiales aux Pays-Bas.

Intitulé *Xenophobic Machines* (« Les machines xénophobes »), ce rapport montre que des critères relevant du profilage racial ont été intégrés lors de l’élaboration du système algorithmique utilisé pour déterminer si des demandes d’allocations familiales devaient être considérées comme erronnées et potentiellement frauduleuses. En conséquence, des dizaines de milliers de parents et de personnes ayant la charge d’enfants, appartenant pour la plupart à des familles à faibles revenus, se sont vu accuser à tort de fraude par les autorités fiscales néerlandaises, les membres de minorités ethniques étant touchés de manière disproportionnée. Si ce scandale a fait tomber le gouvernement néerlandais en janvier, les enseignements qui en ont été tirés restent insuffisants, malgré de nombreuses enquêtes.

Amnesty International

Mesure et réduction des biais en Apprentissage Automatique

Équité dans l’Apprentissage Automatique

- Définitions d’équité
  - Représentation mathématique de différentes notions éthiques
- Métriques d’équité (*fairness metrics)*
  - Mesure du niveau d’équité/de biais d’un modèle ou d’une BDD
- Algorithmes de réduction des biais (*bias mitigation*)

Très grand nombre de définitions, métriques et algorithmes

Définitions et métriques peuvent être contradictoires

Classification équitable (*fair classification*)

- Considère un attribut protégé (ou plusieurs)
- Prédictions indépendantes de l’attribut protégé
- Groupe privilégié vs groupe défavorisé
- Souvent notion de résultat positif et négatif

Exemples de définitions d'équité en classification

Modèle équitable si...

Demographic Parity (DP) :

$$P(\hat{y} = +|G = \text{def}) = P(\hat{y} = +|G = \text{priv})$$

prédiction positive

groupe défavorisé

groupe privilégié

Statistical Parity Difference :

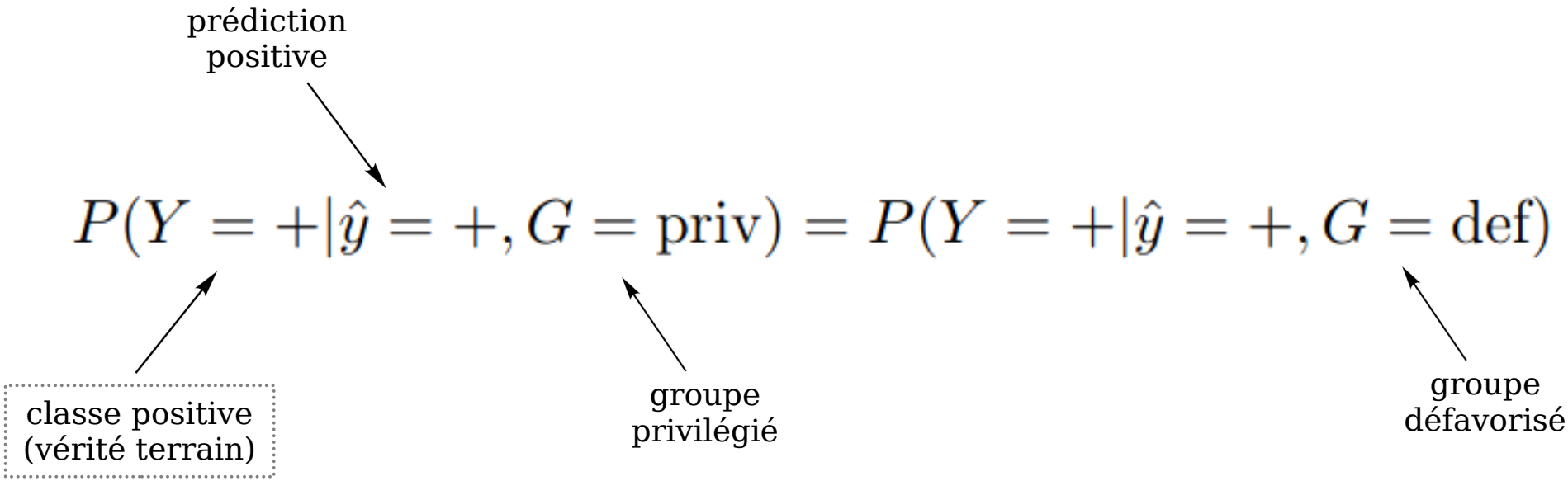
$$bias_{SPD} = P(\hat{y} = +|G = \text{def}) - P(\hat{y} = +|G = \text{priv})$$

Dwork et al, 2013

Exemples de définitions d'équité en classification

Modèle équitable si...

Predictive Parity :

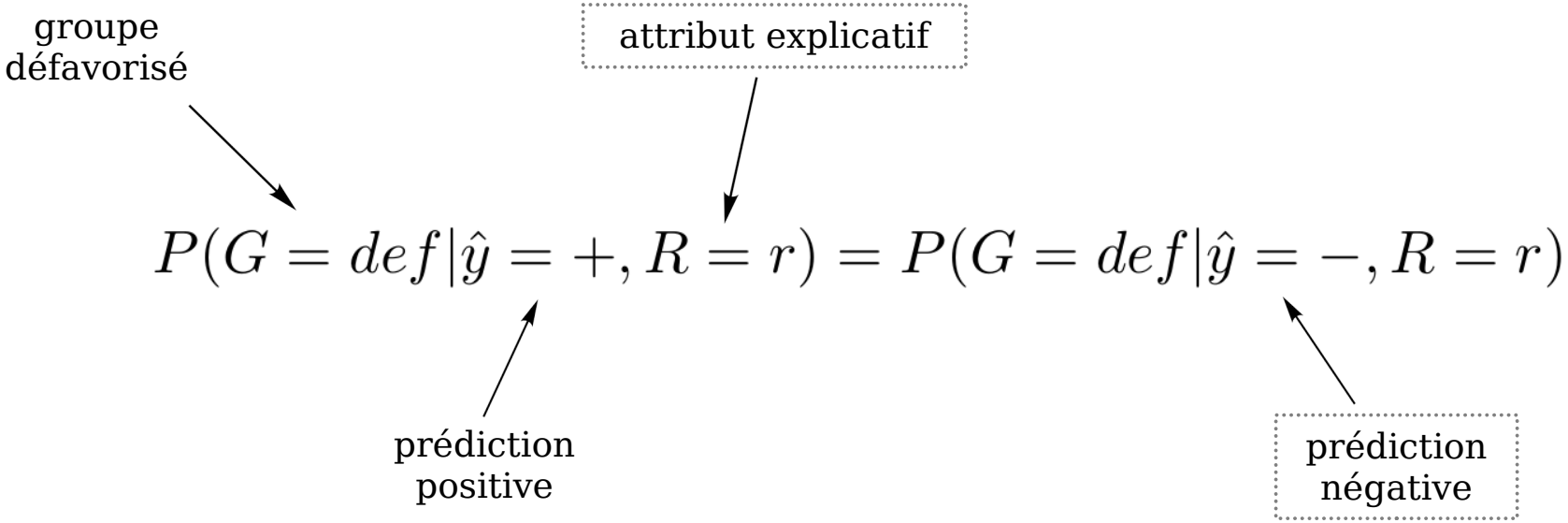


Chouldechova, 2016

Exemples de définitions d'équité en classification

Modèle équitable si...

Conditional Demographic Disparity (CDD)



Wachter et al, 2020

Cadre légal pour l’équité dans l’Union Européenne (UE)

Equité dans l'UE : Un cadre réglementaire qui promeut l'égalité

- Article 14 de la Convention européenne des droits de l'homme : Interdiction de discrimination
- Article 21 de la Charte des droits fondamentaux de l'UE : Non-discrimination

→ Interdiction de discriminer sur base de certains critères protégés

…à moins qu'un objectif légitime, jugé approprié et nécessaire, puisse objectivement et raisonnablement le justifier

Cour de justice de l’Union européenne : Principes légaux clés de la jurisprudence

- **Égalité formelle** : traiter tous les individus de la même manière pour éviter les discriminations
- **Égalité matérielle** : tenir compte du contexte social et des inégalités historiques pré-existantes pour les corriger et atteindre une égalité effective
- **Approche contextuelle** : chaque situation peut être traitée différemment dans différents contextes et en considérant des éléments différents

Lien entre les principes juridiques et les définitions mathématiques d’équité

Égalité formelle et matérielle dans les métriques

- Métrique conservatrice de biais (*bias preserving)*
  - *Reproduit les performances historiques avec un taux d'erreur pour chaque groupe semblable à celui des données d'entraînement*

    Wachter et al, 2021
  - Exemple :
    - Predictive Parity $P(Y = +|\hat{y} = +, G = \text{priv}) = P(Y = +|\hat{y} = +, G = \text{def})$
  - // égalité formelle

Égalité formelle et matérielle dans les métriques

- Métrique transformatrice de biais (*bias transforming)*
  - *Compare les taux de résultats favorables entre les différents groupes et prend en compte les biais sociaux en nécessitant une décision explicite quand aux biais qui devraient être présents dans le système*

    Wachter et al, 2021
  - *Exemples :*
    - Demographic Parity $P(\hat{y} = +|G = \text{def}) = P(\hat{y} = +|G = \text{priv})$
    - Conditional Demographic Disparity $P(G = def|\hat{y} = +, R = r) = P(G = def|\hat{y} = -, R = r)$
  - // égalité matérielle

Approche contextuelle et discrimination explicable

- Attribut conditionnel / explicatif :
  - → Rapprochement de l'approche contextuelle
  - → Autorise une discrimination explicable

  Discrimination totale = discrimination immorale + discrimination explicable

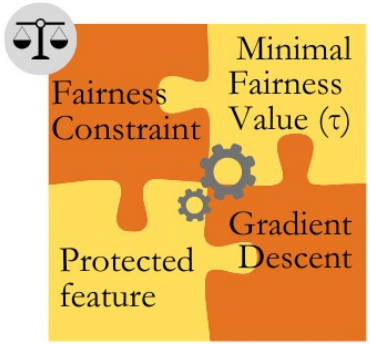
  Kamiran et al, 2013
- Exemples :
  - Conditional Demographic Parity $P(\hat{y} = +|G = \text{def}, R = r) = P(\hat{y} = +|G = \text{priv}, R = r)$
  - Conditional Demographic Disparity $P(G = def|\hat{y} = +, R = r) = P(G = def|\hat{y} = -, R = r)$

Expérimentation : Impact de la définition d’équité sur l’équité des modèles de prédiction

Création et évaluation des modèles avec réduction des biais

Apprendre :

Données d'entrainement



Meta-algorithme
Celis et al, 2019

Modèle de classification plus équitable ?

Évaluer l'équité:

Données de test

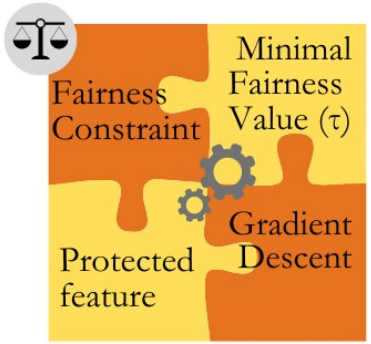
Modèle de classification plus équitable ?

Prédiction plus équitables ?

Création et évaluation des modèles avec réduction des biais

Apprendre :

Données d'entrainement



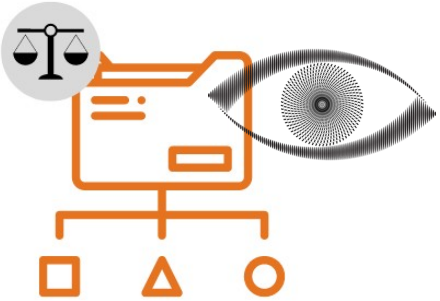
Meta-algorithme
Celis et al, 2019

Modèle de classification plus équitable ?

Évaluer l'équité:

Données de test

Modèle de classification plus équitable ?

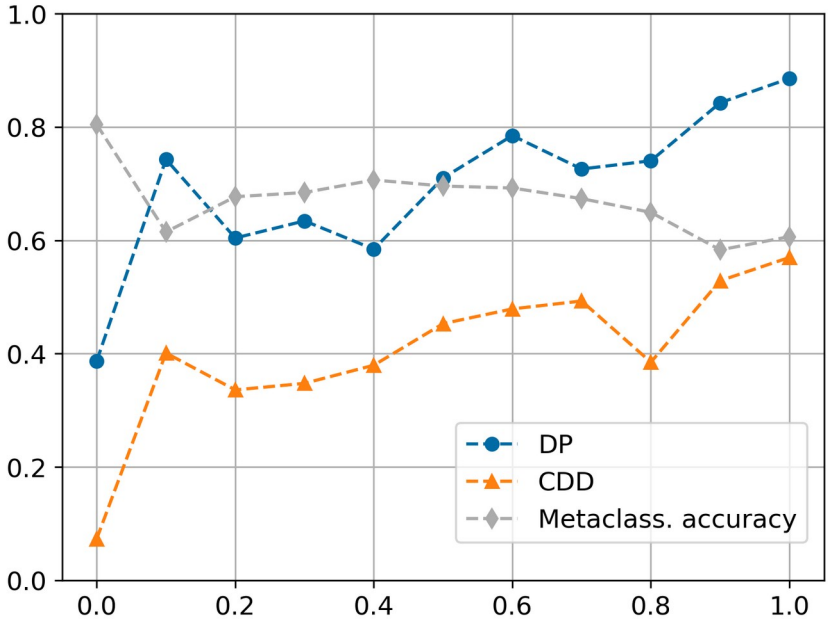


Prédiction plus équitables ?

Résultats préliminaire :
Réduction des biais pour satisfaire Demographic Parity

Adult : Prédiction d'un revenu bas ou haut

Evolution de l'équité des modèles entrainés sur Adult avec la valeur minimale de la contrainte Demographic Parity



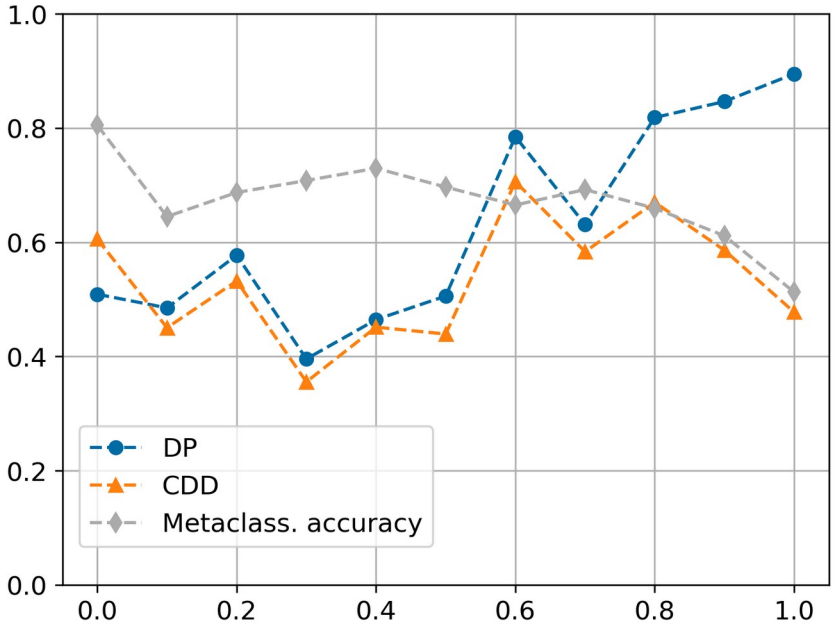
τ, valeur minimale de DP autorisée (force de la contrainte)

Attr. protégé : *race* (blanc & non-blanc)
Attr. explicatif : *education*

- Équité augmente avec la contrainte
- Accuracy diminue avec la contrainte
- Différence entre DP et CDD → discrimination explicable par l'éducation constante
- Correction par rapport à DP ≠ discrimination immorale mesurée par CDD, introduction de nouveau biais ?

Law : Prédiction de l'admission dans les écoles publique de droit

Evolution de l'équité des modèles entrainés sur Law avec la valeur minimale de la contrainte Demographic Parity



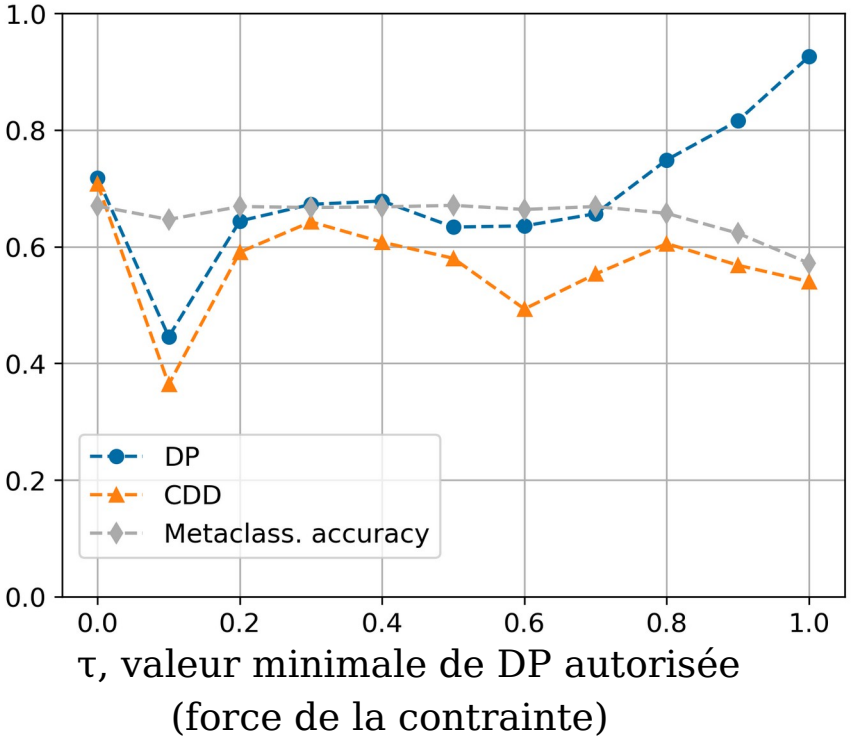
τ, valeur minimale de DP autorisée (force de la contrainte)

Attr. protégé : *race* (blanc & non-blanc)
Attr. explicatif : *GPA*

- Equité selon DP augmente avec la contrainte
- Accuracy diminue avec la contrainte
- Contrainte sur DP corrige (presque) uniquement discrimination immorale jusqu'à τ = 0,7
- Pour les τ = 0,7, correction réintroduit potentiellement de la discrimination injustifiée selon CDD

COMPAS: Prédiction du risque de récidivisme

Evolution de l'équité des modèles entrainés sur COMPAS avec la valeur minimale de la contrainte Demographic Parity



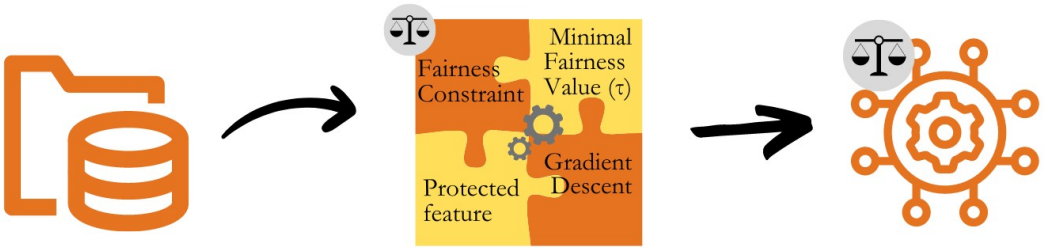
Attr. protégé : *race* (caucasien & noir)
Attr. explicatif : *priors* (nombre de condamnations pénales)
Test set : Arrestation dans les deux ans

- Correction des biais pour $\tau > 0{,}8$
- Différence entre DP et CDD grandit avec la contrainte
- Accuracy varie assez peu
- Réduction du biais utile ?

Discussion des résultats

- Recherche future :
  - Autres définitions d’équité et comparaisons
- Choix du « meilleure » modèle pour accuracy et ≠ notions d’équité ?
  - Résultats différents suivant les scénarios
  - Décision devrait être prise selon le contexte
  - Suggestion d’une obligation légale de justifier ce choix

Métriques d’équité en Apprentissage Automatique et droit de l’Union Européenne en matière de non-discrimination



Corresponding authors :

Magali Legast : magali.legast@uclouvain.be

Lisa Koutsoviti-Koumeri : lisa.koutsoviti@uhasselt.be

Yasaman Yousefi : yasaman.yousefi3@unibo.it

UCLouvain

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

UHASSELT KNOWLEDGE IN ACTION

UNIVERSITÉ DU LUXEMBOURG