# HET METEN VAN DISCRIMINATIE IN ALGORITMISCHE BESLUITVORMING

**Kunstmatige Intelligentie wordt de laatste jaren steeds vaker gebruikt om beslissingen te nemen over keuzes die de levens van mensen beïnvloeden. Dit kan gaan om het bepalen welke online-advertentie het meest geschikt is voor een persoon, of iemand goed zal presteren tijdens een bepaalde baan of wat de kans is dat iemand (nogmaals) een strafbaar feit zal plegen. Een risico is dat deze algoritmes bepaalde groepen benadelen, bijvoorbeeld op basis van geslacht of etniciteit. Is het mogelijk om discriminatie in algoritmische besluitvorming te meten?**

Stan van Loon

Veel mensen zien Kunstmatige Intelligentie (KI) als eerlijk en niet discriminerend. 'Het is een computer die de beslissing maakt, computers kunnen geen fouten maken.' Helaas is dit niet altijd het geval.

Om te begrijpen hoe algoritmes discriminerend kunnen zijn, kijken we eerst naar de manier waarop deze algoritmes gecreëerd worden. Algoritmes worden vaak 'getraind' met behulp van data uit het verleden. Het al-

goritme leert om de redeneringswijze van mensen over te nemen. Op deze manier kan een algoritme, wanneer deze nieuwe data ziet, zelf een keuze maken over de actie die benodigd is bij deze specifieke persoon. Als de data die het algoritme gebruikt discriminerend zijn, zal het algoritme deze discriminatie hoogstwaarschijnlijk overnemen.

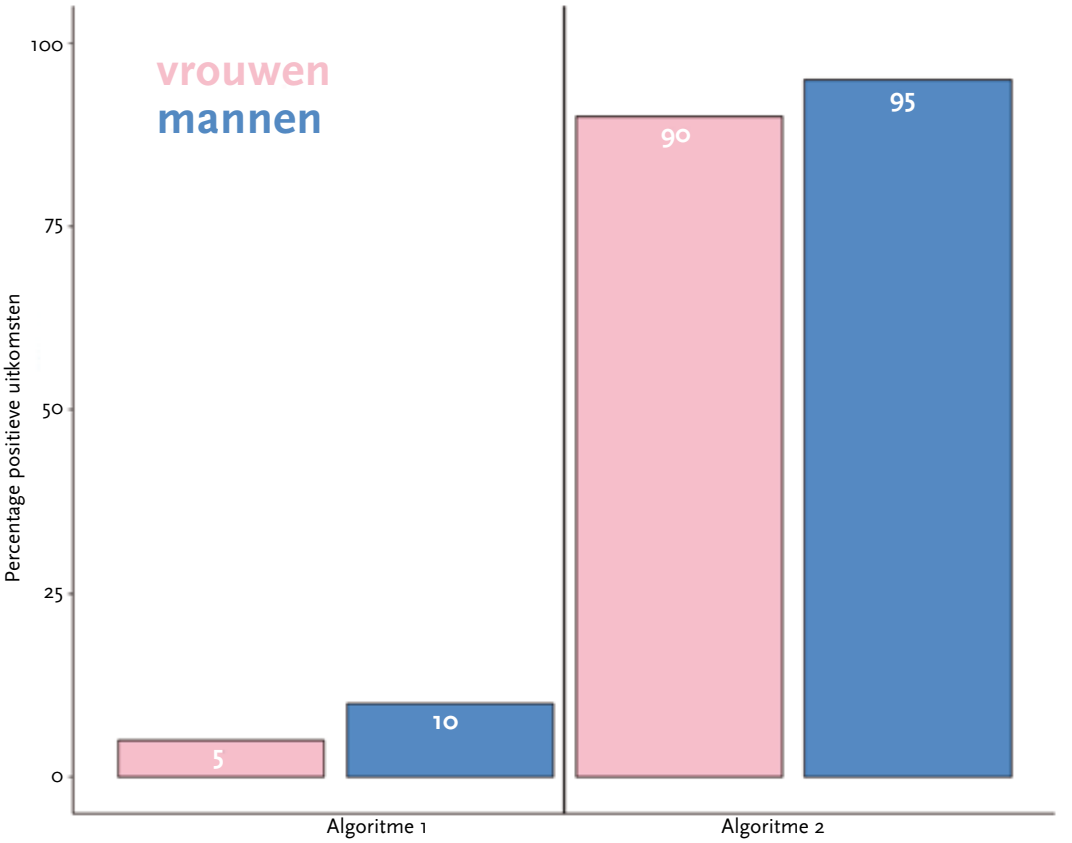
## Amazon

Een goed voorbeeld van een algoritme dat mensen discrimineerde, is een algoritme van Amazon. Amazon wilde haar *recruitment* proces verbeteren door een KI in te zetten om mensen te selecteren voor openstaande vacatures. De meest geschikte mensen werden vervolgens uitgenodigd voor een sollicitatiegesprek. Het algoritme in deze KI werd getraind met data over welke mensen er wel en niet werden aangenomen bij Amazon in de afgelopen tien jaar. In deze periode werden vooral (blanke) mannen aangenomen. Ondanks dat het niet impliciet verteld werd van welk geslacht of etniciteit deze mensen waren, wist het algoritme te discrimineren op geslacht en mogelijk zelfs op etniciteit. Het algoritme leerde op basis van de data die het kreeg, dat het een positieve eigenschap was om een man te zijn. Het kan bijvoorbeeld zo zijn dat het algoritme aan de hobby's van een persoon kan afleiden dat het een man of een vrouw is. Denk aan iemand die voetbal kijken en gamen als hobby's heeft. Deze persoon heeft een veel grotere kans om een man te zijn dan iemand die shoppen als hobby heeft. Het ligt dus vaak niet aan het algoritme of aan de KI zelf, maar aan de data waar het van leert.

Het is dus mogelijk dat Amazon in deze tienjarige periode onbedoeld heeft gediscrimineerd op geslacht en op etniciteit. Amazon probeerde, na het ontdekken van de onbedoelde discriminatie, het algoritme aan te passen om minder discriminerend te zijn. Helaas was het algoritme zo discriminerend dat het niet meer gebruikt kon worden voor het selecteren van mogelijke werknemers. Los van een deukje in de reputatie, leed Amazon geen ernstige gevolgen. Amazon werd niet aangeklaagd voor het discrimineren op basis van geslacht of op basis van etniciteit. (Dastin, 2018)

## Discriminatie in kaart brengen

Aan de hand van deze en andere leermomenten zijn er meerdere onderzoeken uitgevoerd om discriminatie in algoritmes tegen te gaan of te verminderen. In elk van deze onderzoeken werden er bepaalde processen bedacht die discriminatie in algoritmes konden verminderen. Hoeveel discriminatie tegengegaan of verminderd werd, werd vaak berekend met een formule bedacht door een samenwerking tussen de University of Toronto en Microsoft in 2013. Zij bedachten de volgende maatstaf voor het meten van discriminatie in algoritmes: 'De kans dat iemand van de niet-beschermde groep een positieve uitkomst heeft minus de kans dat iemand van de beschermde groep een positieve uitkomst heeft.' In het geval van de recidivistendataset (deze dataset wordt nader besproken) op basis van geslachtsdiscriminatie, kunnen wij dit als volgt interpreteren: 'De kans dat een man geclassificeerd zal worden als een recidivist minus de kans dat een vrouw geclassificeerd zal worden als een recidivist.' Deze maatstaf is alleen van toepassing op de uiteindelijke keuze die een algoritme maakt en kan een eenzijdig beeld geven van de keuzes die een algoritme maakt om tot de beslissing te komen. Bovendien is het mogelijk dat deze maatstaf een verkeerd beeld geeft van de hoeveelheid discriminatie die een algoritme bevat. Dit laten we zien aan de hand van een fictief voorbeeld.




Figuur 1. De kans op een positieve uitkomst voor mannen en vrouwen in twee verschillende algoritmes

In figuur 1 zijn twee verschillende algoritmes te zien. In het eerste algoritme hebben de vrouwen een 5% kans om geclassificeerd te worden als een recidivist en de mannen een 10% kans. In het tweede algoritme zijn deze kansen respectievelijk 90% en 95%. Beide algoritmes zullen als 5% discriminerend beschouwd worden volgens de huidige maatstaf (10 – 5 = 95 – 90 = 5). Echter krijgen mannen een twee keer zo hoge kans op een positieve uitkomst in algoritme 1, in algoritme 2 is deze kans ongeveer 1,06 keer zo hoog. Dit geeft aan dat deze maatstaf een vervormd beeld kan geven van de hoeveelheid discriminatie die zich in een algoritme bevindt. Om dit te voorkomen, is er een methode ontwikkeld die een beter en completer beeld geeft van discriminatie in algoritmes. Aan de hand van een model dat met logistische regressie de kans op recidive schat, worden er vier alternatieve manieren gegeven om discriminatie in algoritmes zichtbaar te maken.

## De dataset

We zullen deze nieuwe methode illustreren aan de hand van de dataset van 'The Florida Department of Corrections' (ProPublica, 1996). In deze dataset staan gegevens over mensen die eenmaal of meerdere malen een strafbaar feit gepleegd hebben. Van deze dataset worden hun leeftijd, aantal eerdere arrestaties, of er geweld plaatsvond in de laatste arrestatie en hoe ernstig de daad is waarvan de persoon beschuldigd wordt gebruikt als variabelen die het algoritme meeneemt in de berekeningen. Aan de hand van deze 'variabelen' bepaalt het algoritme hoe groot de kans is dat elk specifiek persoon nogmaals een strafbaar feit zal plegen, dit is de kans op recidive. Deze berekende kans wordt ook wel de 'modelscore' genoemd.

De data worden gesplitst in train-data om het model op te *fitten* en een test-dataset om de nauwkeurigheid van het model te testen. De maatstaven die gebruikt worden om de graad van discriminatie te laten zien, zijn allemaal metingen op de resultaten van de logistische regressie uitgevoerd op de testdata.

We gebruiken logistische regressie om de kans op recidive te voorspellen en het meten van discriminatie te illustreren. Deze logistische regressie gebruikt de eerdergenoemde variabelen om de kans op recidive te

voorspellen. De formule van de logistische regressie ziet er als volgt uit:

$$\text{Log}\left(\frac{p}{1-p}\right) = 0{,}55 - \text{Graad}\bullet 0{,}31 - \text{Leeftijd}\bullet 0{,}04 + \text{Misdrijven}\bullet 0{,}15 + \text{Geweld}\bullet 3{,}24 + \text{Geslacht}\bullet 0{,}21$$

| | | |
|---|---|---|
| $p$ | = | De kans op recidive. |
| Graad | = | De graad van het misdrijf. Dit is een getal van 0 tot 11. Hoe groter dit getal, hoe ernstiger het misdrijf. |
| Misdrijven | = | Aantal eerder gepleegde misdrijven. |
| Geweld | = | 0, geen geweld bij het misdrijf. 1, anders. |
| Geslacht | = | 0, als Vrouw. 1, als Man. |

## Discriminatie weergeven in de modelscores

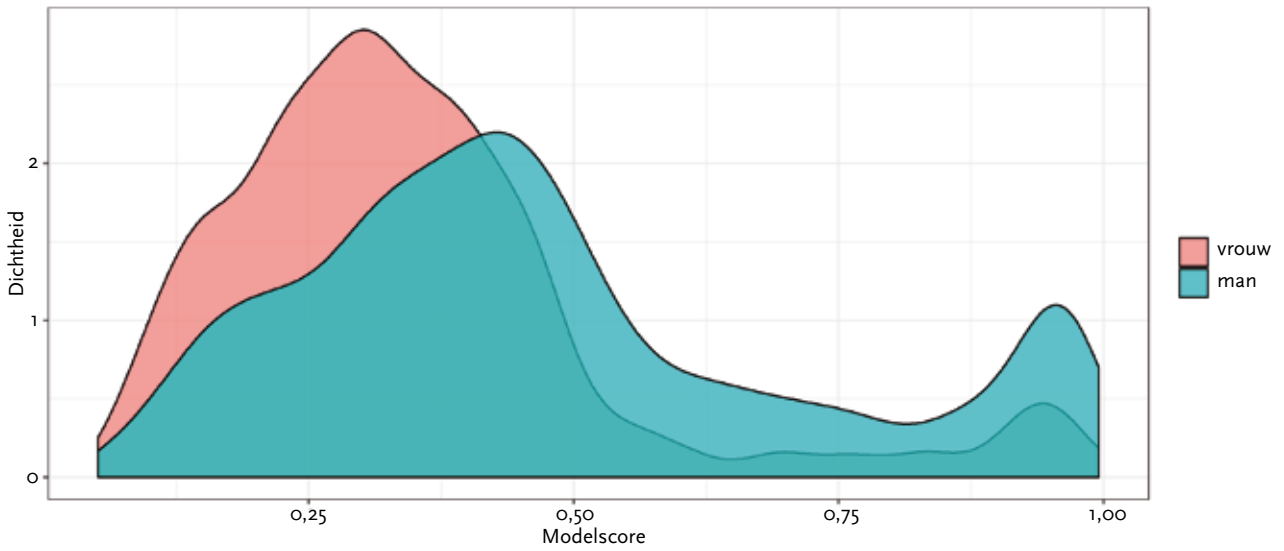
Om discriminatie op basis van geslacht weer te geven kunnen de modelscore-verdelingen van mannen en vrouwen met elkaar vergeleken worden. Dit wordt gedaan in figuur 2. Hier is te zien dat een logistische regressie (die geslacht ook daadwerkelijk als een voorspellende variabelen meeneemt) over het algemeen mannen een hogere modelscore geeft. Dit betekent dat de logistische regressie mannen sneller als recidivist zal classificeren dan vrouwen. Dit komt doordat er later een cutoff gezet wordt voor het bepalen of iemand als mogelijke recidivist bestempeld wordt. Hoe dit precies werkt wordt later in dit artikel uitgelegd.

Het is natuurlijk mogelijk dat de twee verdelingen erg op elkaar lijken en elkaar grotendeels overlappen. Als dit het geval is kan de Kolmogorov-Smirnov test gebruikt worden om te testen of de twee verdelingen significant van elkaar verschillen. In dit geval heeft de Kolmogorov-Smirnov test een p-waarde lager dan $2{,}2 \times 10^{-16}$, een p-waarde die nauwelijks van nul te onderscheiden is. Dit geeft aan dat de twee verdelingen hoogstwaarschijnlijk van elkaar verschillen en dat er onderscheid gemaakt wordt tussen mannen en vrouwen in deze logistische regressie.

## $AUC_D$

Om op een makkelijker te begrijpen maatstaf voor discriminatie te komen, kunnen de modelscores van de mannen en de vrouwen met elkaar vergeleken worden en worden omgezet in een getal. Dit wordt gedaan met behulp van de Area Under the Curve of AUC in het kort. De AUC wordt oorspronkelijk gebruikt als een maatstaf om de nauwkeurigheid van een model te bepalen. De AUC geeft weer hoe groot de kans is dat een willekeurig persoon die uiteindelijk een recidivist blijkt te zijn een hogere modelscore heeft gekregen dan iemand die uit eindelijk geen recidivist blijkt te zijn.

In dit geval wordt de AUC gebruikt als een maatstaf voor discriminatie, vanaf nu wordt dit de $AUC_D$ genoemd. De $AUC_D$ geeft de kans weer dat een willekeurige man een hogere modelscore heeft dan een willekeurig gekozen vrouw. In het geval dat de $AUC_D$ gelijk is aan 1, heeft elke man een hogere modelscore dan elke vrouw. In het geval dat de $AUC_D$ gelijk is aan 0, geeft dat aan dat elke vrouw een hogere modelscore heeft dan elke man. Als de $AUC_D$ berekend wordt over figuur 2 komt hier een waarde uit van ongeveer 0,69. Wat aangeeft dat een willekeurige man uit de testdata een 69% kans heeft om een hogere




Figuur 2. De modelscore-verdelingen van een logistische regressie van de recidivisten-dataset

modelscore te krijgen dan een willekeurige vrouw uit de testdata.

Op deze manier kan de discriminatie in de modelscores van een algoritme al weergegeven worden met een enkel getal.

## Discriminatie weergeven in de keuze

Nadat de modelscores voor elk individu berekend zijn kan er een *cutoff* bepaald worden. Deze cutoff bepaalt of een specifieke waarde van elke modelscore gelabeld wordt als 'positief' of als 'negatief'. In het geval van de recidivisten-dataset worden alle individuen van de testdata met een modelscore gelijk of lager dan 0,5 geclassificeerd als 'niet-recidivisten' en alle individuen in de testdata met een modelscore boven de 0,5 worden gelabeld als een recidivist. Uiteindelijk kan er gekeken worden naar de spreiding van recidivisten en niet-recidivisten en van mannen en vrouwen. Om deze spreiding weer te geven kan een kruistabel opgezet worden. Een voorbeeld van een kruistabel is te zien in tabel 1.

| | | VOORSPELLING | |
|---|---|---|---|
| | | Niet-recidivist | Recidivist |
| GESLACHT | Vrouw | *298* | *46* |
| | Man | *955* | *505* |

Tabel 1. Kruistabel van de keuze van de logistische regressie op de testdata

Aan deze kruistabel is te zien dat mannen relatief vaker worden geclassificeerd als recidivisten volgens de logistische regressie. Als we dit percentage van de mannen vergelijken met het daadwerkelijke percentage van de mannen uit de dataset, valt op dat deze aanzienlijk lager is: 0,35 volgens het algoritme en 0,47 in de dataset. Voor de vrouwen zijn deze percentages respectievelijk 0,13 en 0,36. Over het algemeen classificeert de logistische regressie mensen snel als niet-recidivisten, dit heeft te maken met de vaste cutoff bij een score van 0,50. In figuur 2 valt op dat er veel waarnemingen waren net onder de 0,50. Dit kan verklaren waarom er bij deze cutoff minder mensen geclassificeerd worden als recidivist. Hierdoor hebben vrouwen over het algemeen een kans van maar liefst 13% om geclassificeerd te worden als recidivist. Dit algoritme toont dus sterke positieve discriminatie tegenover vrouwen. Mogelijk wordt dit veroorzaakt door het meenemen van het geslacht in de logistische regressie. Wat zal er gebeuren als deze variabele niet wordt gebruikt?

Helaas, zelfs als de variabele van het geslacht uit de logistische regressie gehaald wordt, blijkt de logistische regressie nog steeds discriminerend te zijn ten opzichte van de feitelijke waarnemingen. Volgens de regressie hebben mannen een 32% kans en vrouwen een 19% kans om geclassificeerd te worden als een recidivist. Deze kansen liggen al dichter bij elkaar en liggen ook dichter bij daadwerkelijke waarnemingen, maar laten nog steeds een duidelijk nadeel zien voor mannen. Dit heeft te maken met de andere variabelen die de logistische regressie gebruikt om de kans op recidive te voorspellen. Aan de hand van deze variabelen is het mogelijk om te voorspellen of een persoon een man of vrouw is. Dit gebruikt het algoritme om alsnog een onderscheid te kunnen maken tussen mannen en vrouwen.

Een andere manier om discriminatie in een getal uit te drukken is volgens de *selection rate*. De selection rate in dit geval geeft de verhouding aan tussen de percentages van mannen en vrouwen die geclassificeerd worden als recidivisten. De berekening voor de selection rate gaat als volgt:

$$\frac{505/(955+505)}{46/(298+46)} \approx 2{,}5$$

Deze 2,5 geeft weer dat mannen, na berekening door de logistische regressie, een 2,5 keer zo grote kans hebben om als een recidivist geclassificeerd te worden ten opzichte van vrouwen. Om te kijken of hier sprake is van onterechte discriminatie, gebruiken we de 80%-regel om te bepalen of er een kans is dat er onethisch gediscrimineerd wordt. Deze regel wordt door de overheid van Amerika gehanteerd om te bepalen of er mogelijk sprake is van discriminatie. Volgens de 80%-regel mag dit getal variëren tussen de 0,80 en de 1,25. Alles wat buiten dit gebied valt, kan mogelijk gezien worden als discriminatie. Het is natuurlijk mogelijk dat dit niet het geval is. Deze maatstaf is alleen een indicatie die aangeeft of verder onderzoek nodig is. (Workplace Dynamics, LLC, 2009)

Als laatste maatstaf kan er een Chi-kwadraat test gedaan worden over de kruistabel. Deze test geeft aan of er een verband is tussen de twee variabelen. In dit

geval zijn dat de beslissing van de logistische regressie en het geslacht. Als uit de Chi-kwadraat test blijkt dat er een verband is, zou dit in aanmerking komen voor discriminatie. In dit geval geeft ook de Chi-kwadraat test een p-waarde die vrijwel gelijk is aan nul en dit duidt op mogelijke discriminatie binnen het algoritme.

Met deze methode kunnen niet alleen logistische regressies, maar vele soorten algoritmes onderzocht worden op mogelijke discriminatie, zoals neurale netwerken en random forests. Deze methode richt zich niet alleen op discriminatie op basis van geslacht, maar is toepasbaar op bijvoorbeeld rassendiscriminatie en op leeftijdsdiscriminatie.

Samenvattend hebben we vier alternatieven gegeven om discriminatie door algoritmes inzichtelijk te maken. Door middel van een grafiek van de modelscore verdelingen, twee statistische toetsen, de $AUC_D$ en de 80%-regel, kan er een compleet beeld gecreëerd worden van de hoeveelheid discriminatie. De resultaten van de maatstaven van één model kunnen dan weer vergeleken worden met die van een ander model om te kijken of één van de modellen minder discriminerend is dan de andere.

Literatuur


Dataset: Broward County Clerk's Office, Broward County Sheriff's Office, Florida Department of Corrections, ProPublica. (2019). COMPAS Recidivism Risk Score Data and Analysis. ProPublica Data Store. https://bit.ly/STAtOR2qolTjU

Calmon, F., Wei, D., & Vinzamuri, B., Varshney, K., (2017). *Optimized Pre-processing for Discrimination Prevention*. https://bit.ly/STAtOR35WiBTY

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. https://bit.ly/STAtOR2rXmYzG

Pedreschi, D., Ruggieri, S., & Turini, F., (2009). *Measuring Discrimination in Socially-Sensitive Decision Records.* http://pages.di.unipi.it/ruggieri/Papers/sdm09.pdf

Pedreschi, D., Ruggieri, S., & Turini, F., (2012). *A Study of Top-K Measures for Discrimination Discovery*. http://pages.di.unipi.it/ruggieri/Papers/sac2012.pdf

Workplace Dynamics, LLC., (2009). *Adverse Impact Analysis; Four-Fifths Rule*. https://bit.ly/STAtOR2PaE3hn

Zemel, R., Wu, Y., Swerky, K., & Pitassi, T., (2013). *Learning Fair Representations*. https://bit.ly/STAtOR2Rk40Ok



Stan van Loon is in 2019 afgestudeerd bij de studie Toegepaste Wiskunde op de Hogeschool van Amsterdam. Tijdens zijn afstudeerstage deed hij, in samenwerking met Pegasystems, onderzoek naar een methode om discriminatie in algoritmes meetbaar te maken. Dit artikel is een samenvatting van het afstudeeronderzoek. Het hele onderzoek is te vinden via https://bit.ly/STAtOR2P19V7Y
E-mail: stanvanloon1998@gmail.com



## Causality & Complexity

**Het bestuur van de VVSOR hoopt op 12 maart 2020 veel leden en andere belangstellenden te ontmoeten op de VVSOR Annual Meeting in de Gertrudiskapel – In De Driehoek congres- en vergadercentrum in Utrecht. Het thema van de dag is Causality and Complexity. Verdere bijzonderheden kunt u vinden in de volgende editie van STAtOR en te zijner tijd op de website van de VVSOR. Zet de datum van 12 maart alvast in uw agenda.**