Samenvatting

Onderzoek kwantitatieve methoden om algoritmische discriminatie te meten en te voorkomen.

College voor
de Rechten van de Mens

Inleiding

Organisaties maken steeds vaker gebruik van algoritmes. Door algoritmische selecties en categorisering proberen organisaties werkprocessen efficiënter te maken. Algoritmegebruik brengt echter risico's mee op discriminatie.

In de computerwetenschap zijn de afgelopen jaren verschillende fairness-methodes ontwikkeld: kwantitatieve methodes voor het meten en voorkomen van discriminatie door algoritmes. Dat is een positieve ontwikkeling, maar de vraag is of dergelijke methodes afdoende zijn om discriminatie in juridische zin te kunnen toetsen en voorkomen. Het College heeft hiernaar onderzoek laten doen door de Technische Universiteit Eindhoven.

Belangrijkste bevindingen

Het onderzoek concludeert dat *fairness*-methodes overeenkomsten bevatten met de beoordelingskaders die worden gebruikt om vast te stellen of er sprake is van discriminatie. Deze kwantitatieve methodes alleen zijn echter niet voldoende zijn om discriminatie aan te tonen of te voorkomen.

Het toetsen en voorkomen van discriminatie in algoritmes vereist multidisciplinaire samenwerking, actieve monitoring en verankering van het toetsen op discriminatie in de werkprocessen.

Fairness-methodes kunnen wel onderdeel zijn van de toets waarmee bepaald wordt of er sprake is van discriminatie. De rest van deze samenvatting geeft handvatten hoe organisaties *fairness*-methodes in kunnen zetten in lijn met het recht op gelijke behandeling.

Handvatten

Voor het aanpakken van algoritmische discriminatie met *fairness*-methodes

Directe discriminatie vaststellen en voorkomen

Discriminatie is het ongelijk behandelen, achterstellen of uitsluiten van mensen op basis van bepaalde persoonlijke kenmerken (de discriminatiegronden [1]), zonder dat daar een rechtvaardiging voor is. Met discriminatie wordt onderscheid tussen groepen mensen gemaakt. Dit kan direct of indirect gebeuren.

Bij algoritmes is sprake van directe discriminatie als een variabele, categorie of selectie-instrument expliciet refereert aan een beschermde groep. Het gebruiken van de variabele 'geslacht' zorgt bijvoorbeeld voor direct onderscheid op grond van geslacht en zal dus ook leiden tot onderscheid tussen mannen en vrouwen in een algoritme. Dit is niet toegestaan, tenzij er een wettelijke uitzondering is. Directe discriminatie in de context van algoritmes vindt bijna altijd plaats als een criterium van het algoritme ook een discriminatie-grond is, zoals geslacht of nationaliteit. Directe discriminatie kan dus voorkomen worden door criteria die ook discriminatiegronden zijn niet mee te nemen in het algoritme.

Directe discriminatie in CV-matching

Een voorbeeld van directe discriminatie is het gebruik van een CV-matchingtool om te filteren voor een technische functie. Het algoritme bevat de variabele geslacht, waarbij het kandidaten die zich identificeren als vrouw of non-binair automatisch lager scoort of zelfs direct afwijst. Omdat de selectie op basis van het beschermde persoonskenmerk plaatsvindt, geslacht, is er sprake van directe discriminatie.

1 Discriminatie op de volgende gronden is wettelijk niet toegestaan: godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, seksuele gerichtheid, burgerlijke staat, handicap of chronische ziekte, leeftijd, arbeidsduur, vast of tijdelijk contract en WAZO-verlof.

Toetsen van indirecte discriminatie met *fairness*-methodes

Of een algoritme indirect onderscheid maakt, is lastiger om vast te stellen. Het gaat dan namelijk om op het oog neutrale variabelen. Het *effect* van op het oog neutrale variabelen kan zijn dat sommige groepen daardoor relatief vaker geselecteerd worden. Zo'n effect noemen we *indirect onderscheid*. Als er geen zwaarwegende, goede reden ('objectieve rechtvaardiging') bestaat om zo'n variabele te gebruiken, is sprake van indirecte discriminatie.

Het is belangrijk om als organisatie in beeld te krijgen waar indirect onderscheid door algoritmes plaatsvindt, en hoe sterk dit effect is, zodat dit ook opgelost kan worden. Er zijn meerdere mogelijkheden om indirect onderscheid vast te stellen. **Eén van de manieren is statistisch in kaart te brengen of variabelen een ongelijk effect hebben op beschermde groepen, bijvoorbeeld met *fairness*-methodes.** Veelgebruikte methodes die verschillen tussen (beschermde) groepen in kaart brengen heten *group fairness*-methodes.

Voorbeelden van dit soort methodes zijn het berekenen van selectieratio's en misclassificatierisico's voor verschillende groepen. Selectieratio's kwantificeren of beschermde groepen vaker of minder vaak geselecteerd worden. Misclassificatieratio's geven weer wie naar verhouding vaker onterecht wel of niet geselecteerd wordt. wordt. Dat wordt gedaan door te kijken naar valspositieven[2] en valsnegatieven.[3]

Indirect onderscheid via postcode-variabele in fraudemodellen

Een voorbeeld van indirecte discriminatie is het gebruik van een bepaald model om fraude op te sporen. De organisatie neemt postcode op als factor, omdat men ervan uitgaat dat in sommige postcodegebieden vaker fraude voorkomt dan in andere. Het model bevat niet de variabele migratieachtergrond en maakt dus geen direct onderscheid. Maar omdat mensen met verschillende etnische achtergronden in Nederland niet evenredig verspreid zijn, maar vaak geconcentreerd wonen over postcodegebieden, maakt zo'n variabele waarschijnlijk indirect onderscheid op grond van etnische herkomst.

2 Valspositieven zijn voorbeelden waarvoor het algoritme een positief label voorspelt maar het daadwerkelijke label negatief is.
3 Valsnegatieven zijn voorbeelden waarvoor het algoritme een negatief label voorspelt maar het daadwerkelijke label positief is.

Vier vragen waarmee een organisatie na kan gaan of er sprake is van indirecte discriminatie in een algoritme met *group fairness* methodes

1. Welke uitkomsten moeten worden vergeleken? Wat betekent nadeel in de context van het algoritme?

Het hangt van de context van het algoritme af of het een voordeel of nadeel is om geselecteerd te worden. Bijvoorbeeld: een CV-matchingtool selecteert relatief meer mannelijke dan vrouwelijke sollicitanten. Niet geselecteerd worden is dan een nadeel. Maar in de context van fraudecontroles is geselecteerd worden een nadeel.

2. Welke (beschermde) persoonskenmerken worden meegenomen?

Bepaal welke (wettelijk beschermde) groepen mogelijk benadeeld worden door het algoritme. Met andere woorden: welke discriminatiegronden zijn relevant in de context van algoritme? Een veelvoorkomende uitdaging is een gebrek aan toegang tot variabelen die beschermde persoonskenmerken meten.

Verwerking van bijzondere persoonsgegevens

De verwerking van (bijzondere) persoonsgegevens is niet zomaar toegestaan. Er is doorgaans een wettelijke grondslag nodig. Soms kunnen uitzonderingen in de AVG gebruikt worden om persoonsgegevens te kunnen gebruiken bij onderzoek naar discriminatie. In sommige gevallen kan via het CBS statistisch onderzoek worden uitgevoerd naar verhoudingen een specifieke populatie, zonder dat de uitkomsten herleid kunnen worden naar individuen (pseudonimisatie). Het is aan te raden bij de Autoriteit Persoonsgegevens te informeren wat wel en niet is toegestaan.

3. Wat is de juiste referentiepopulatie?

Om discriminatie vast te stellen, moet vastgesteld worden dat een groep ongelijk behandeld is in een gelijk geval. Bepaal hiervoor welke groep er in een vergelijkbare situatie verkeert maar niet benadeeld wordt. Dit wordt mogelijk ook bepaald door (het gebrek aan) data dat beschikbaar en geschikt is.
Om bijvoorbeeld vast te stellen of een fraudecontrolesysteem voor midden- en kleinbedrijven onderscheid maakt op grond van herkomst van de eigenaar, is de referentiepopulatie andere bedrijven uit diezelfde sector van belang, niet alle midden- en kleinbedrijven in Nederland.

4. Bij welke grenswaarde moet (niet) worden ingegrepen?

In de praktijk zal een selectieratio of misclassificatieratio zelden (precies) gelijk zijn voor verschillende groepen. Er bestaat geen algemeen toepasbare grenswaarde of methode om te bepalen of er sprake van onderscheid is, dit moet altijd per geval worden bekeken. Bij onderscheid, of een vermoeden daarvan, moeten de stappen van de objectieve rechtvaardiging van de indirecte discriminatietoets doorlopen worden om na te gaan of er sprake is van discriminatie. Loop hiervoor de volgende vragen af:

- Streeft het algoritme een legitiem doel na?
- Zijn de variabelen die onderscheid maken geschikt om het gestelde legitieme doel van het algoritme te bereiken?
- Zijn de variabelen die onderscheid maken noodzakelijk om het gestelde legitieme doel te bereiken?
- Zijn de variabelen die onderscheid maken, alles afwegend, proportioneel? Met andere woorden: staan alle nadelen van het gemaakte onderscheid in redelijke verhouding tot de naar verwachting te bereiken doelen?

Vijf uitgangspunten om te bepalen of er sprake is van discriminatie:

1. Het is belangrijk om deze analyse uit te voeren in een multidisciplinair team, met in ieder geval een juridisch expert, domeinexpert en een datawetenschapper.
2. De context is van belang. Discriminatie kan niet in een simpel cijfer of ja/nee antwoord gevat worden.
3. Leg eerst in simpele taal uit wat het nadeel precies is en definieer het probleem in eenvoudige taal (na de analyse). Vertaal het daarna pas naar kwantitatieve (fairness-)metriek.
4. Het is niet altijd mogelijk om statistisch volledig betrouwbaar te zijn, met name door (te) kleine datasets of een groot aantal categorieën te analyseren. Bij statistische onzekerheid is de stelregel 'bij twijfel niet inhalen' (de variabele niet gebruiken) wat het College betreft het devies.
5. Het bepalen van de referentiepopulatie, de groep in een vergelijkbare situatie, om na te gaan of er onderscheid is, kan lastig zijn. Dit hangt ook weer samen met beschikbare data. Probeer een balans te vinden tussen beschikbare data en welke groep op papier het best dient om te vergelijken.

Is er onderscheid gevonden waarvoor geen objectieve rechtvaardiging is? Dan moet hiervoor een oplossing worden gezocht.
Om dit goed te doen, is het belangrijk om te weten wat de onderliggende oorzaak van het gevonden onderscheid is. Er zijn grofweg twee soorten oorzaken die het risico op discriminatie in algoritmes kunnen vergroten: verschillen in voorspelkracht en een statistisch verband tussen een discriminatiegrond en een doelvariabele.

Als een algoritme minder accuraat is voor bepaalde (beschermde) groepen dan heeft het algoritme een lage voorspelkracht voor deze groepen. Dit kan leiden tot discriminatie. Een lage voorspelkracht kan veroorzaakt worden door gebrekkige datakwaliteit of ontwerpkeuzes. In 2022 oordeelde het College bijvoorbeeld dat de inzet van antispieksoftware voor het surveilleren van tentamens kan leiden tot indirect onderscheid op grond van ras als de software slechter presteert voor mensen met een donkere huidskleur.

Datgene wat een algoritme probeert te voorspellen heet een doelvariabele. Als er een statistisch verband is tussen de doelvariabele en een discriminatiegrond, dan zal een accuraat algoritme dit verband reproduceren. Dit kan leiden tot onderscheid. Statistische verbanden kunnen komen door een bestaande ongelijkheid in de samenleving die het algoritme reproduceert maar ook door keuzes in dataverzameling. Dit kan misgaan als een doelvariabele niet goed gedefinieerd is, bijvoorbeeld als het gebaseerd is op bestaande vooroordelen of omdat de verzamelde data niet representatief is. Discriminatie op de arbeidsmarkt komt veelvuldig voor. Een algoritme voor werving en selectie van nieuwe werknemers kan deze bestaande vooroordelen, zoals ongelijke kansen van vrouwelijke werknemers, reproduceren.

NB: het kan zijn dat er geen oplossing gevonden kan worden voor een (ongerechtvaardigde) overselectie van beschermde groepen door een algoritme. In dat geval mag het algoritme niet gebruikt worden.

Focus ook op voorkomen van discriminatie in algoritmes

Naast het toetsen van effecten bij algoritmes die al in gebruik zijn, is het belangrijk om waar mogelijk te focussen op het voorkomen van discriminatie.

In de computerwetenschap zijn veel methodes ontwikkeld om onderscheid aan te pakken, waarbij het probleem wordt behandeld als een optimaliseringsprobleem: een zo hoog mogelijke voorspelkracht van het algoritme behalen onder een randvoorwaarde die mogelijk onderscheid kwantificeert. In de praktijk zijn deze methodes vaak ineffectief en kan discriminatie zelfs worden verergerd, omdat deze methodes veelal geen rekening houden met de oorzaak van het onderscheid.

De oorzaak snappen is cruciaal om preventieve maatregelen te kunnen nemen. Bijvoorbeeld: als verschillen tussen groepen komen door te weinig (goed voorspellende) data, dan zijn maatregelen tijdens het trainen van het model maar beperkt effectief.

Statistiek om algoritmische discriminatie beoordelen in de rechtspraak

In de rechtspraak wordt nog relatief weinig gebruik gemaakt van statistische methodes om onderscheid vast te stellen. Om meer inzicht te krijgen in de statistische methodes die in juridische procedures worden gebruikt, heeft onderzoeksbureau Enabling Digital Rights in opdracht van het College eerder onderzoek uitgevoerd over de toepassing van statistische methodes in de rechtspraak. Uit dit onderzoek blijkt dat in juridische procedures statistiek soms wordt toegepast om discriminatie vast te stellen. statistiek soms wordt toegepast om discriminatie vast te stellen in juridische procedures. Er blijkt ook uit dat de rechtspraak geen eenduidige methodes hanteert voor het meten van onderscheid.