BEGELEIDENDE PRESENTATIE – BIAS ANALYSE

Uitvoeren van een bias analyse over ontwikkelde algoritmes

CPA vergadering 10 december 2020

CTO Innovatieteam Gemeente Amsterdam:

Data Scientist: [redacted]
Privacy jurist: [redacted]
Product Owner: [redacted]

Adviezen ingewonnen bij Privacy Officers:
[redacted]

Inleiding

- Binnen de gemeente Amsterdam worden in toenemende mate data gedreven, digitale producten gerealiseerd, bv. het vakantieverhuur algoritme;
- Algoritmes worden veelal ingezet met als doel om met een risico indicatie de beschikbare capaciteit effectiever en efficiënter in te zetten;
- Uit jurisprudentie en de bestaande praktijk blijkt dat de inzet van (zelflerende) algoritmes binnen de overheid van meerwaarde kan zijn, mits is voldaan aan wettelijke vereisten;
- In lijn met de AVG en de Ethics Guideline for Trustworthy AI van de Europese Commissie, dienen algoritmes **wettig, ethisch en robuust** te zijn en moet worden voldaan aan beginselen als transparantie, uitlegbaarheid en controleerbaarheid;
- De Kindertoeslagaffaire en de uitspraak inzake SyRI hebben tot extra inzichten geleid om nauwkeuriger en diepgaander te onderzoeken of ontwikkelde algoritmes de juridische en ethische toets kunnen doorstaan;
- De afgelopen maanden is in de Werkplaats gewerkt aan een methodiek om te kunnen controleren of de onderliggende data en het ontwikkelde algoritme een ongerechtvaardigd verschil maken tussen personen en groepen.



Proces


Verkennen business behoefte en problematiek
Diepgaande analyse proces en beschikbare data
Ophalen en toetsen bestaande business rules uit werkproces
Vaststellen juridische waarborgen en te verwerken data
Onderzoek- en ontwikkelfase algoritme
Uitvoeren Bias Analyse
Beoordelen 'features'
Integrale toetsing aan AI Europese normen
Inwinnen advies CPA, FG
Nemen collegebesluit en communicatieplan
Opstellen validatieplan en uitvoeren Pilot
Informeren gemeenteraad
Evalueren bevindingen
Leren, verbeteren en doorontwikkelen

Bias analyse en weging features



Onderzoeken



Analyseren bevindingen, toetsing en correctie



Gewenst eindresultaat

- Onderzoeken of de onderliggende data waarop het algoritme is ontwikkeld een bias bevat;
- Onderzoeken welke features in het algoritme een bias bevatten.

- De dialoog voeren en toetsen of er al dan niet sprake is van een ongerechtvaardigde aanpak in de uitvoering van het werkproces en feature in het algoritme;
- De dialoog voeren en toetsen of de betreffende features al dan niet in het algoritme blijven of in hoeverre andere maatregelen kunnen worden getroffen om het negatieve effect te beperken;
- Noodzakelijke aanpassingen en verbeteringen worden doorgevoerd binnen het algoritme en het werkproces.

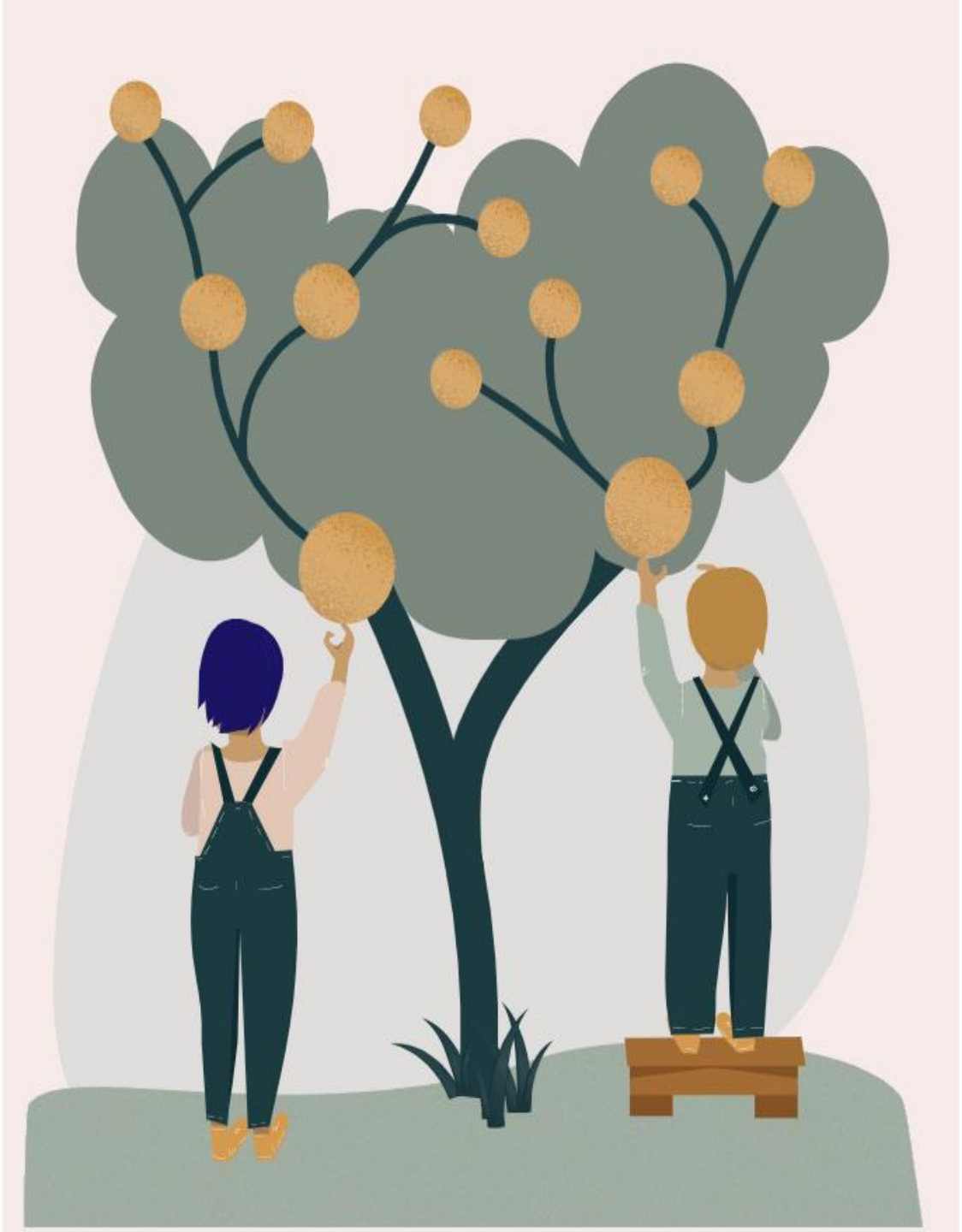
- Het algoritme is getoetst aan het Europese normenkader en is **niet discriminerend**;
- Het algoritme is in vergaande mate transparant, verifieerbaar en controleerbaar.

Wat is bias?

"Wanneer een model uitkomsten produceert die, op wat voor manier dan ook, ongerechtvaardigd zijn"*

Veel verschijningsvormen:

- Bepaalde groepen worden er vaker ten onrechte uitgepikt voor nader onderzoek
- Voor bepaalde groepen is de pakkans lager
- De score van een individu verandert als diegene tot een andere groep had behoord (terwijl die groep irrelevant zou moeten zijn)



*We zijn nog zoekende naar een goede, Nederlandse term en definitie.

Terminologie (1)

Eenheid

Datgene wat we onderzoeken

- Een Nederlander
- Een adres in Amsterdam
- Een bijstandsontvanger in Amsterdam



Populatie

Alle eenheden waarvan we mogelijk data zouden kunnen hebben

- Alle Nederlanders
- Alle adressen in Amsterdam
- Alle bijstandsontvangers in Amsterdam



Sample (ook wel: steekproef)

De eenheden waarvan we daadwerkelijk data hebben

- Nederlanders die een bepaalde vragenlijst hebben ingevuld
- Adressen in Amsterdam die onderzocht zijn op vakantieverhuur
- Bijstandsontvangers die in onderzoek zijn geweest



Terminologie (2)

Target (ook wel: doelvariabele)

Datgene wat voorspeld wordt

- Risico op diabetes
- Risico op illegale vakantieverhuur
- Risico op onrechtmatigheid van bijstandsuitkering



Label (ook wel: uitkomst)

De target variabele van een specifieke eenheid

- Mark heeft geen diabetes (label = 0)
- Op Amstel 1 vindt geen illegale vakantieverhuur plaats (label = 0)
- De uitkering van Mark is onrechtmatig (label = 1)



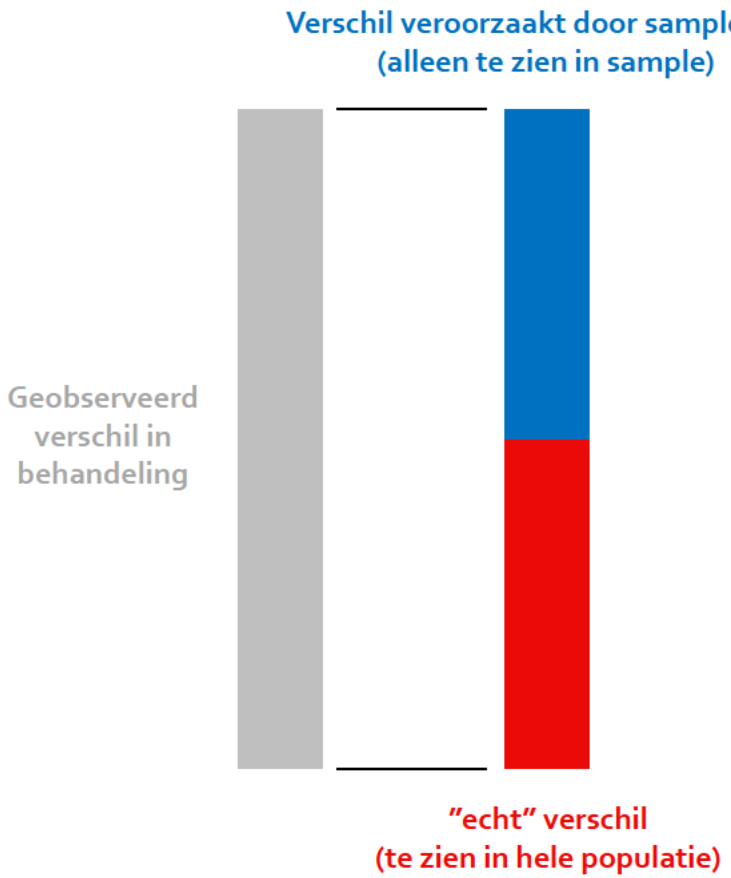
Feature (ook wel: variabele)

Gegeven over een eenheid, waarmee een voorspelling gemaakt wordt

- Mark sport 2x per week
- Amstel 1 heeft geen woonbestemming
- Mark is kortgeleden verhuisd



Sample bias als oorzaak van verschillen



Sample bias
Wanneer de sample afwijkt van de populatie (bijvoorbeeld in eigenschappen of uitkomsten)

Soorten sample bias*



Selection bias: Sommige eenheden hebben een lagere of juist hogere kans om in de sample terecht te komen

- Effect: De voorspellingen van het model weerspiegelen niet de patronen uit de hele populatie; het model presteert slechter voor ondergerepresenteerde groepen.
- Voorbeelden: In bepaalde buurten wordt meer gesurveilleerd door de politie, waardoor voor die buurten meer criminaliteit geregistreerd wordt; door vooroordelen worden er eerder meldingen bij de gemeente ingediend tegen mensen met een migratie-achtergrond.



Self-selection bias: Eenheden bepalen zelf of ze in de sample terecht komen (speciaal geval van *selection bias*)

- Effect: Zie *selection bias*.
- Voorbeeld: Mensen met geldproblemen zijn minder snel geneigd om een vragenlijst over hun financiële situatie in te vullen.



"Besmette" labels: Labels uit het verleden kloppen niet

- Effect: Het model gaat de "besmette" uitkomsten herproduceren.
- Voorbeeld: Sollicitatieprocedure was in het verleden niet objectief, waardoor het model vrouwen slechter beoordeelt.

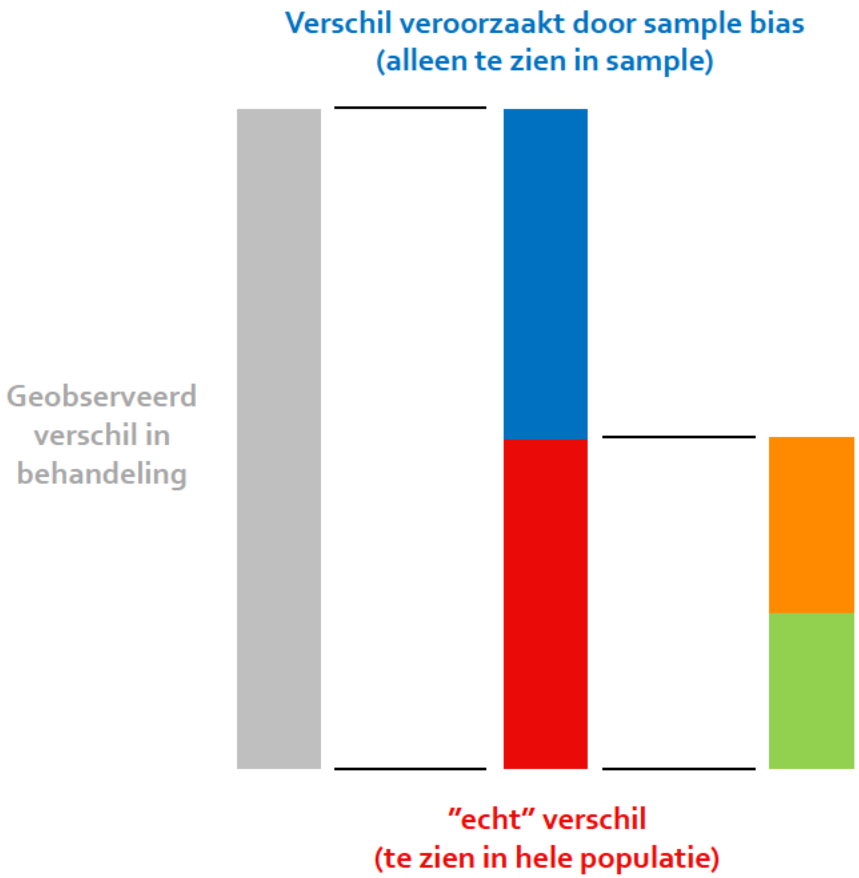


Beperkte features: Sommige features zijn minder informatief of betrouwbaar voor een deel van de populatie

- Effect: Het model presteert slechter voor een bepaalde groep.
- Voorbeeld: Iemand die moeite heeft met de taal maakt fouten bij het invullen van een formulier.

*Dit zijn een aantal veelvoorkomende soorten bias, maar deze lijst is zeker niet alomvattend.

Aanvaardbare en niet-aanvaardbare verschillen



Sample bias
Wanneer de sample afwijkt van de populatie (bijvoorbeeld in eigenschappen of uitkomsten)

Vraagstukken

- We hebben doorgaans geen gegevens van de hele populatie, dus hoe kunnen we de **sample bias** onderscheiden van **echte verschillen**?
- Vinden we echte verschillen **aanvaardbaar**, of ligt er iets aan ten grondslag waardoor ze ethisch **onaanvaardbaar** zijn (bijv. historische of huidige ongelijkheid in de samenleving)?

*Bijvoorbeeld: Een universiteit selecteert van de vrouwelijke aanmeldingen een lager percentage dan van de mannelijke aanmeldingen, maar dit komt doordat vrouwen zich vaker aanmelden voor studies met veel concurrentie.

Oorsprong van verschillen kunnen we slechts deels achterhalen



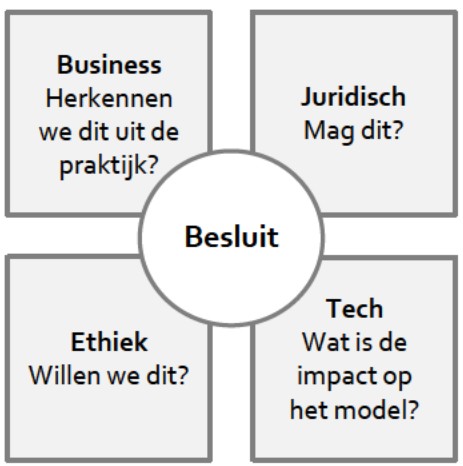
Hoe kunnen we de **sample bias** onderscheiden van **echte verschillen**?

- Vergelijk sample met een volledig willekeurige steekproef uit de populatie (labels + features)
  - Voorbeeld: Onderzoek 300 willekeurige adressen in Amsterdam op illegale vakantieverhuur
- Vergelijk sample met bekende gegevens op populatieniveau (alleen features)
  - Voorbeeld: Voor sommige features bestaan statistieken van het CBS



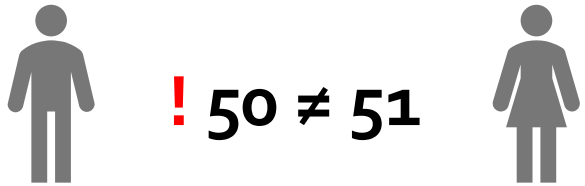
Zijn echte verschillen **aanvaardbaar** of **onaanvaardbaar**?

- Logisch redeneren over mogelijke achterliggende redenen van een verschil
- Afweging maken tussen waarde voor het model en ethische wenselijkheid



Hoe meet je bias?

Kernprincipe: Vergelijk de uitkomsten van het model voor twee groepen of individuen met elkaar



Wie vergelijken we met elkaar?

- Noodzaak om “gevoelige groepen” te definiëren

Wat vergelijken we precies?

- Dat hangt er vanaf wat we belangrijk vinden: Wat vinden we “eerlijk”?

Gevoelige groepen

"Wie kunnen we vergelijken met elkaar?"

- Nationaliteit*
- Geboorteland*
- Huidskleur
- Etniciteit
- Geslacht*
- Leeftijd*
- Burgerlijke staat*
- Seksuele geaardheid
- Religie
- Politieke opvatting
- Verblijfsstatus*
- Zwangerschap
- Gezondheid (handicap, ziekte)
- Sociale klasse
- Genetica

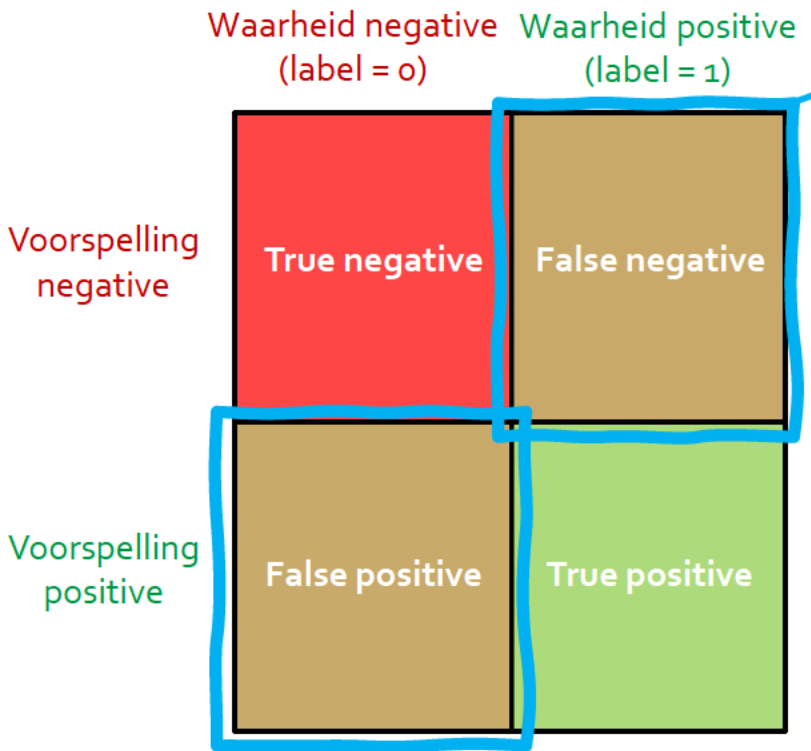
*Data beschikbaar binnen de gemeente

Gewenste gegevens voor het uitvoeren bias analyse

- Nationaliteit
- Geboorteland
- Geslacht
- Leeftijd
- Burgerlijke staat
- Verblijfsstatus

Model kan op verschillende manieren de fout in gaan

"**Wat** vergelijken we met elkaar?"



Voor bepaalde groepen is de pakkans lager

True positive & true negative Het model had het goed

False positive & false negative Het model zat ernaast

False positive Een eenheid wordt ten onrechte geselecteerd

- Voorbeeld: Een bijstandsuitkering wordt aangemerkt als onrechtmatig, maar is dit niet.
- Gevolg: Iemand wordt ten onrechte onderzocht.

False negative Een eenheid wordt ten onrechte *niet* geselecteerd

- Voorbeeld: Een bijstandsuitkering wordt aangemerkt als rechtmatig, maar is dit niet.
- Gevolg: Er glipt een fraudegeval door de selectie heen.

Bepaalde groepen worden er vaker ten onrechte uitgepikt voor nader onderzoek

Wat te doen als een gevoelig kenmerk niet beschikbaar is?



Sommige gevoelige kenmerken (bijvoorbeeld huidskleur, religie) worden niet geregistreerd*.

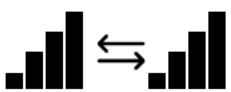
Als we niet weten wie tot welke groep behoort, kunnen we deze groepen ook niet vergelijken.



Oplossing

… werk in uitvoering

Idee: Schat deze kenmerken bij benadering op basis van sociologische en statistische kennis over de groepen.



Bijvoorbeeld: Ouders met een niet-westerse migratieachtergrond hebben gemiddeld meer kinderen dan ouders zonder migratieachtergrond.

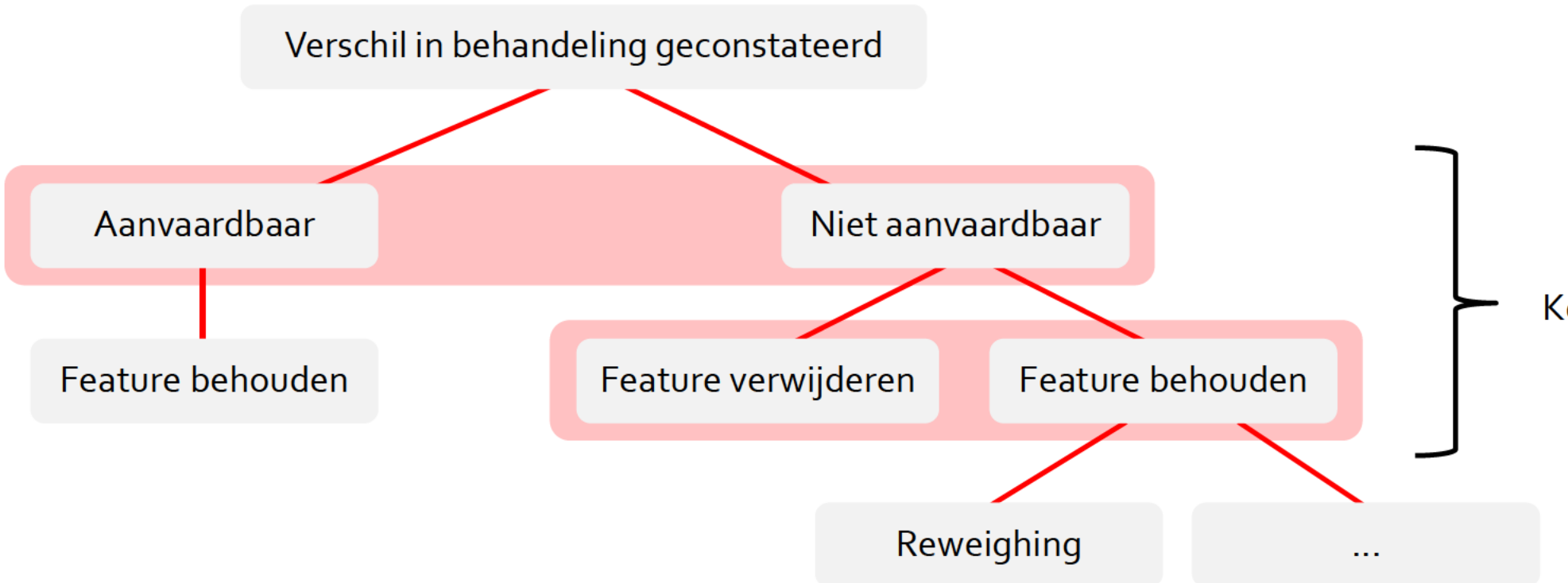
Mogelijk samenwerking met ketenpartners als het CBS?

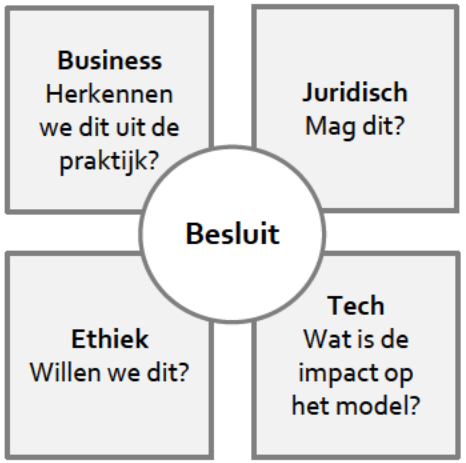


Zal niet mogelijk zijn voor alle gevoelige kenmerken.

*In veel gevallen met goede reden

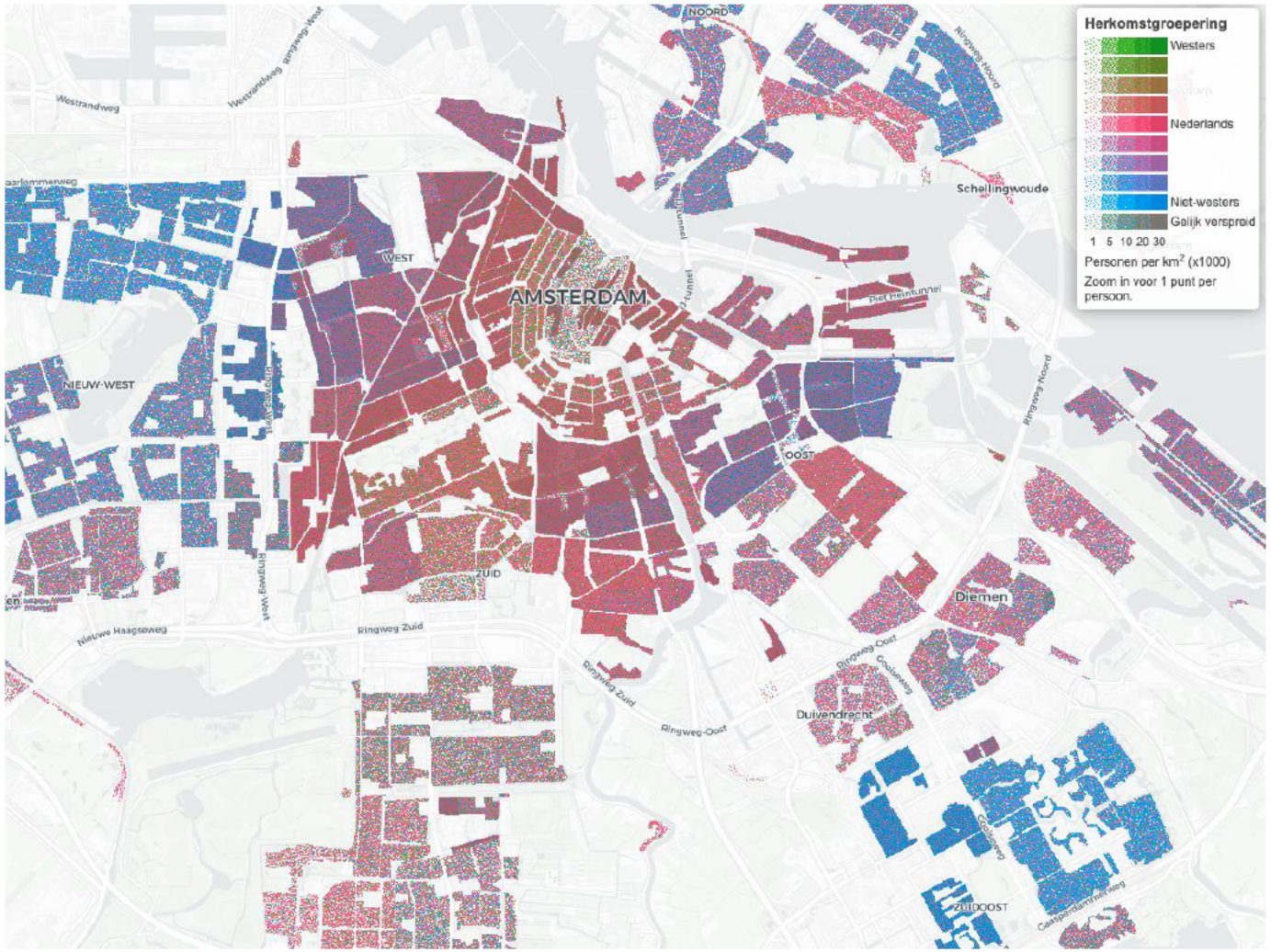
Verschil geconstateerd, wat dan?





Proces van vele iteraties!

Wat gebeurt er als je postcode meeneemt in een model?



Alleen gevoelige features weglaten is vaak niet genoeg!

Bijvoorbeeld: zelfs als nationaliteit niet in het model zit, kan door postcode onderscheid gemaakt worden op nationaliteit.

Screenshot van:
http://research.cbs.nl/colordotmap/NL/index.html

Ongewenste features weglaten is vaak niet genoeg door indirecte bias

Directe bias		**Indirecte bias**
Direct effect van de ongewenste feature op de uitkomsten van het model	**vs.**	Indirect effect van de ongewenst feature op de uitkomsten van het model, omdat beide samenhangen met een derde variabele (de **proxy**)

Dus na het verwijderen van een feature, moeten we altijd nog een ronde draaien en de bias opnieuw meten en beoordelen.

Om feature te behouden, kunnen we het model of de data aanpassen

Een optie is *reweighing*:

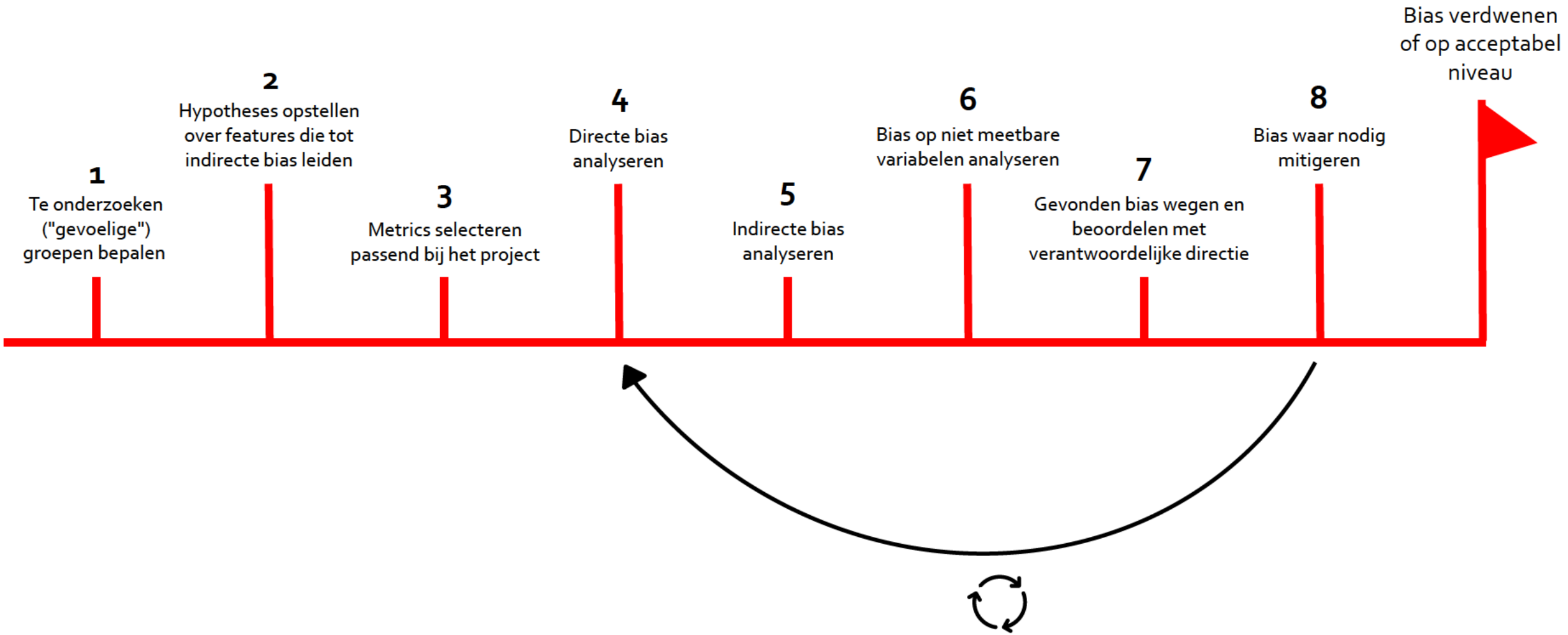
- Verminder bias in de dataset **vóór het trainen** van een model (preprocessing) door **gewichten** toe te kennen aan iedere rij in de dataset.
- Stel: Data laat zien dat mannen met grotere waarschijnlijkheid een positieve uitkomst ontvangen.

Sex	Highest degree	Job type	Positive job application	Weight
M	H. School	Board	+	0.75
M	University	Board	+	0.75
M	H. School	Healthcare	+	0.75
M	University	Healthcare	-	2
M	H. School	Board	+	0.75
F	H. School	Education	-	0.67
F	University	Education	-	0.67
F	H. School	Board	+	1.5
F	None	Healthcare	+	1.5
F	University	Education	-	0.67

- *Male positives* en *female negatives* krijgen een **laag gewicht,** zodat het lijkt alsof er hiervan minder in de dataset zitten.
- *Female positives* en *male negatives* krijgen een **hoog gewicht,** zodat het lijkt alsof er hiervan meer in de dataset zitten.

Voorbeeld uit "Data preprocessing techniques for classification without discrimination" (Kamiran & Calders, 2012)

Stappenplan



Toetsingskader voor ontwikkelde features

Lineair uitgewerkt, feitelijk iteratief proces

1. Is begrijpelijk voor de medewerkers van het betreffende werkproces wat wordt bedoeld met de features in het algoritme?
2. Zijn deze features passend bij de uitvoering van het betreffende werkproces?
3. Wat is het effect van het toepassen van elke feature in het algoritme en de som der delen?
4. In hoeverre is een feature van belang voor het model?
5. Kunnen bepaalde features als risicovol worden beschouwd, omdat ze iets zeggen over persoonlijke kenmerken van personen of groepen?
6. Zijn er features die gebiasedzijn en ervoor zorgen dat bepaalde groepen, bv. mannen of vrouwen, uiteindelijk anders worden behandeld door het algoritme?



Toetsingskader voor ontwikkelde features

Lineair uitgewerkt, feitelijk iteratief proces

7. Om te bepalen of een feature (in)direct discriminerend is moet worden getoetst aan de volgende cumulatieve criteria (behoorlijkheidstoets):

- de **tegenover elkaar gestelde gevallen** zijn in relevante opzichten **voldoende vergelijkbaar**;
- er is **onderscheid** gemaakt tussen deze gevallen;
- het onderscheid heeft geleid tot **nadeel** in behandeling;
- het onderscheid is **redelijk** en **objectief gerechtvaardigd** (generalistische argumenten zijn onvoldoende):
    - *het doel van de discriminerende handelswijze, maatregel of middel dient te beantwoorden aan een noodzakelijke doelstelling van sociaal beleid van een Lid-staat. Dit is het vereiste van legitimiteit;*
    - *de handelwijze, maatregel of middel moet geschikt zijn om dat doel te bereiken. Dit betreft het vereiste van doelmatigheid;*
    - *de handelwijze, maatregel of middel moet daarvoor ook noodzakelijk zijn. Dit is het vereiste van proportionaliteit.*

8. Na voorgaande toetsing zal de eindverantwoordelijke een definitieve keuze moeten maken over het behouden, verwijderen of toepassen van aanvullende maatregelen om de features te gebruiken. Er kunnen vervolgens adviezen worden ingewonnen en een besluit worden genomen door het college.



Benodigde team

Vereiste expertise voor de beoordeling van de features (eerstelijns toets):

- Senior medewerkers werkproces
- Senior jurist werkproces
- Beleidsmedewerker
- Privacy officer
- Data scientist (ontwikkelaar algoritme)
- Statisticus
- Data ethicus

Akkoord op werking model:

- Verantwoordelijk afdelingshoofd / directeur en/of MT gezamenlijk

Inwinnen adviezen bij (tweedelijns toets)

- Security Officer
- Privacy Officer
- CPA
- Functionaris Gegevensbescherming
- Burgerpanel (nog uit te werken)

Besluit in gebruik nemen algoritme (derdelijns toets):

- College

