Advies bias analyse over ontwikkelde algoritmen 10 december 2020

Uittreksel van het verslag van het openbare agendapunt van de vergadering op donderdag 10 december 2020. Uitvoeren bias analyse over de ontwikkelde algoritmes (Cluster Ruimte & Economie).

De voorzitter merkt op dat een uitvoerige beschrijving van de plannen vooraf is toegestuurd (zie vergaderstukken). Hij geeft de betrokkenen de gelegenheid die nu toe te lichten.

Er volgt een presentatie die als bijlage bij dit verslag is opgenomen.
Actuele ontwikkelingen – de toeslagenaffaire, het SyRI-vonnis – hebben duidelijk gemaakt dat het nodig is ontwikkelde algoritmes grondig te onderzoeken om te bezien of zij de juridische en ethische toets kunnen doorstaan. In de Werkplaats is daarom gewerkt aan een methodiek om te kunnen controleren of de onderliggende data en het ontwikkelde algoritme een ongerechtvaardigd verschil maken tussen personen en groepen. Er is een methodiek ontwikkeld om algoritmes te analyseren op ingebouwde vooringenomenheid (bias), met een bijbehorend werkproces dat stapsgewijs beschreven is als een cyclus van verkenning en analyse van data tot en met evaluatie en voorstellen voor verbetering in een volgende fase. Het is dus een iteratief proces dat wordt teruggekoppeld met de betreffende inhoudelijke vakgroep en beleidmakers. Het gewenst eindresultaat is een niet-discriminerend algoritme dat transparant, verifieer- en controleerbaar is. Uiteindelijk moet het een gemeentelijke standaard voor algoritmes opleveren.

In de presentatie wordt vervolgens uitgelegd wat in dit kader wordt verstaan onder bias ('Wanneer een model uitkomsten produceert die, op wat voor manier dan ook ongerechtvaardigd zijn') en wordt de terminologie van het model uitgelegd (eenheid, populatie, steekproef, target/doelvariabele, label/uitkomst en feature/variabele).

Bij de steekproef kan een bias optreden als die afwijkt van de samenstelling van de totale populatie ten gevolge van de selectiecriteria, zelfselectie, besmette features die niet (meer) overeenstemmen met de actuele stand van zaken of beperkte features die minder of minder betrouwbare informatie opleveren over een deel van de populatie. De vraag die dan gesteld moet worden of een bias leidt tot aanvaardbare of onaanvaardbare verschillen. Belangrijk is dan wel of te achterhalen valt hoe de betreffende variabele verdeeld is in de totale populatie. Als bijvoorbeeld bekend is dat vrouwen zich vaker inschrijven voor studies waar veel belangstelling voor is en uit het steekproefonderzoek komt naar voren dat meer vrouwen bij de selectie afgewezen worden, dan is dat een gerechtvaardigd verschil in de onderzoeksuitkomst.
Om te achterhalen of een bias te maken heeft met de steekproef, kan men vergelijkingen maken met andere willekeurige steekproeven uit de populatie, bijvoorbeeld via het CBS of de GBA. Om te bepalen of een verschil gerechtvaardigd is volgt men logische redeneringen waarbij historische achtergronden, praktijkervaringen, ethische en juridische kwesties en onderzoekstechnische aspecten de revue passeren.

Een bias komt men op het spoor door de uitkomsten van het model voor verschillende groepen met elkaar te vergelijken. Vervolgens speelt de vraag of het gaat om gevoelige groepen en belangrijke 'eerlijke' verschillen. Er zijn lijsten opgesteld met gevoelige classificaties van groepen en gegevens die daarover bij de gemeente bekend zijn om de bias analyse te kunnen uitvoeren. Als over een bepaald kenmerk geen gegevens beschikbaar zijn, is er geen vergelijking tussen

eenheden te maken en de vraag is of je dat dan wel zou willen. Het model kan op twee manieren tot foute bevindingen leiden: de voorspelling is positief terwijl de werkelijke uitkomst negatief is (wat betekent dat een eenheid ten onrechte geselecteerd is voor nader onderzoek) of de voorspelling is negatief terwijl de uitkomst positief is (wat betekent dat een eenheid ten onrechte niet geselecteerd is en dus in de praktijk een kleiner pakkans heeft).

Als er verschillen geconstateerd worden, moet bepaald worden of die aanvaardbaar zijn of niet. Vervolgens kan men besluiten de betreffende feature uit het model te verwijderen of wegen te zoeken om die wel te behouden maar de bias tegen te gaan door aanpassingen in het model of verschillende gewichten toe te kennen aan de onderdelen van de dataset. Die beslissing wordt in overleg met de inhoudelijke afdeling genomen. Daarbij moet men onderlinge samenhangen niet uit het oog verliezen. Men kan bijvoorbeeld het attribuut nationaliteit weglaten, maar de postcode geeft dan in de uitkomsten alsnog een indicatie. Er is een stappenplan in gebruik dat loopt van het onderzoek om gevoelige groepen te bepalen en het formuleren van hypotheses over features die tot indirecte bias leiden, analyse van directe en indirecte bias en het wegen en mitigeren van aangetroffen bias tot het wegwerken of op een acceptabel niveau brengen van de bias.

De CPA is benieuwd wat deze analyse oplevert voor het eerder hier gepresenteerde onderzoek naar illegaal vakantieverhuur.

Dat onderzoek is met deze methode getoetst op een aantal features. Een belangrijke was bijvoorbeeld het geslacht van de hoofdbewoner. Mannen bleken vaker aan vakantieverhuur te doen dan vrouwen, maar dat is verklaarbaar op historische gronden omdat mannen op dit moment nu eenmaal vaker hoofdbewoner zijn. Zij liepen dus al een hoger risico in het onderzoek meegenomen te worden. De feature geslacht zat in de dataset omdat men moest weten op welke aanspreektitel men hoofdbewoners moest aanschrijven. In de praktijk bleek het voor de vakafdeling niet relevant of er bij een melding een man of vrouw in het geding was. Voor de uitkomsten van het onderzoek maakte het ook geen verschil als deze feature weggelaten werd. Het is dus van belang goed na te denken over de data die men in een onderzoek opneemt. Als die niets toevoegen aan de voorspellende waarde, maar gezien historische omstandigheden wel een rol spelen, moet de vraag gesteld worden of opname in het algoritme gerechtvaardigd is. Dat vraagt een proces van toetsen en afwegen.

De CPA is benieuwd of er ook een wetenschappelijke validatie is voor dit analysemodel en of er samenwerking plaatsvindt met een universiteit of academische onderzoekers.
Het is pionieren, maar op basis van wetenschappelijke literatuur op het gebied van datascience. Er zijn nog weinig praktijkvoorbeelden, maar er is wel veel theorie beschikbaar. Samenwerking vindt plaats met TADA (UVA) en DEDA(Utrecht). Er komt een rapport van de VNG met standaarden die in ontwikkeling zijn. Het is een voortdurend denkproces en men komt gaandeweg uit bij de kern van de wettelijke taken, waarbij deze aanpak kan leiden tot herijking van werkprocessen.

De CPA werpt de vraag op of er niet een Amsterdams kader nodig is om de vakafdelingen te laten meebewegen in deze ontwikkeling.

De CPA waardeert het dat nu enkele criteria voor gevoeligheid zijn geïdentificeerd, maar denkt dat veel algoritmes daarnaast nog andere (verborgen) onderscheidingen herbergen, zoals arm-rijk. De vraag is hoe men daar goed mee om kan gaan en of het niet aan te bevelen zou zijn om te gaan werken met divers samengestelde teams van ervaringsdeskundigen om blinde vlekken op te sporen.

Het gaat dan vaak om zaken die in de gemeentelijke databanken niet staan geregistreerd, zoals sociale klasse en sociaaleconomische factoren. Of het kan gaan om onmeetbare verborgen variabelen. Aan de samenstelling van de teams worden wel eisen gesteld, maar nog niet op het gebied van diversiteit. De suggestie om te gaan werken met divers samengestelde panels van ervaringsdeskundigen wordt meegenomen.

De CPA krijgt de indruk dat men hier de techniek van via datamining zoeken naar onvermoede correlaties een richting probeert te geven zonder de ongewenste neveneffecten. De vraag is of deze methode toereikend zal zijn om als overheid publieke taken uit te oefenen en te handhaven. Deze methode van opsporen zal niet leiden tot daadwerkelijk optreden. De vraag is bovendien of men met deze aanpak voorkomt dat steeds in dezelfde cirkel gezocht wordt, zoals in de psychiatrie voornamelijk onderzoek naar biologische factoren plaatsvindt en haast niet meer naar psychische of sociale, waardoor de beschikbare data voor nadere analyse beperkt zijn tot wat onder die lantaarnpaal gevonden is. De paradox van big-data-onderzoek is dat men altijd wel een correlatie vindt, maar dat er geen garantie is dat men beter weet hoe zaken echt in elkaar steken.

Met behulp van het algoritme kan men bepalen welke zaken verder onderzoek verdienen. Dat nader onderzoek doorloopt dan een aantal voorgeschreven stappen om te bepalen of er een zaak is waar de overheid een handhavingstaak is. Het algoritme zelf zal dat nooit kunnen bepalen. Dat blijft een menselijke beslissing aan de hand van vastgelegde strafprocedures en overwegingen van proportionaliteit.
Met deze aanpak wil men juist proberen de filtering aan de voorkant zo eerlijk mogelijk te maken zodat niet steeds dezelfde groep uit de algoritmes tevoorschijn komt. Het gaat weliswaar om grote databestanden, maar het zijn mensen die in deze aanpak beredeneerd selecties maken voor de analyses en zo willen voorkomen dat het zoeklicht zich slechts op een beperkte kring richt. Achter de schermen worden veel gesprekken gevoerd om deze aanpak niet te beperken tot non-discriminatie, maar ook andere sociale rechten daarbij te betrekken. Dit alles om te komen tot evenwichtigere bestuursorganen die met een wijze blik zaken in ogenschouw nemen. Dat kan gaan betekenen dat men terug moet naar de tekentafel om met elkaar te bepalen welk soort data wezenlijk zijn om te verzamelen.

De CPA is benieuwd of er geen spanningsveld gaat optreden als men enerzijds transparantie wil bieden en anderzijds handhaving en pakkansen wil ver beteren. Ontstaat dan niet het risico dat kwaadwillenden daarop gaan inspelen en dat wellicht heel nieuwe vooroordelen ontstaan?

Deze aanpak maakt expliciet wat impliciet al gebeurt. Op een gegeven moment zal men zich moeten verstaan met beleidmakers wat hiervan in beleidsregels vertaald moet worden zodat Amsterdammers weten waar ze aan toe zijn. Eerst gaat men kijken wat er overblijft als deze analyse is uitgevoerd op dit algoritme. Dan komt de vraag wat en tot welk detailniveau hiervan moet worden opgenomen in het algoritmeregister. Daar zal men elkaar ergens in moeten vinden, maar daar is nog geen definitief antwoord op.

De voorzitter merkt op dat in het strafrecht zich een vergelijkbaar proces voltrekt. Zodra mensen weten op welke zaken speciaal gelet wordt, houden ze daar rekening mee in hun gedrag.

De CPA merkt op dat het aantal mensen dat in een woning verblijft ook een verborgen indicator voor nationaliteit kan zijn en dus een bias zou kunnen veroorzaken. Verder is de vraag waarom de postcode niet opgenomen is in de lijst van kenmerken, al was het maar omdat men daardoor

verplaatsingen van verschijnselen door de stad kan volgen. Het advies is verder om ook de indicatoren met lage risico's in de analyse te blijven betrekken. Verder vestigt de CPA de aandacht op de bias die het gevolg kan zijn van de capaciteit aan mensen in het werkproces die gaandeweg hun eigen ervaring en voorkeuren inclusief mogelijke vooringenomenheid opbouwen bij de selectie van te onderzoeken situaties. Verder is men benieuwd of voorzien is in een audit voor dit model en adviseert men periodiek het model te testen met gestratificeerde steekproeven om het actueel te houden.

De postcode is niet opgenomen in de lijst omdat het daar gaat om bijzondere persoonsgegevens waarvoor de gemeente een gedegen grondslag moet geven om die te mogen verzamelen. In de analyse worden echter wel alle onderdelen getest op mogelijke (indirecte) bias, ook de postcode. Er wordt ook over gedacht om in de toekomst naast black labels ook te gaan werken met white labels. Over de inrichting van een audit wordt achter de schermen al nagedacht en ook over vormen om het model actueel te houden. Het model wordt ontwikkeld in een brede koker en zal nog meermalen in een of andere vorm bij de CPA passeren.

De voorzitter bedankt voor de goede presentatie van een interessante theoretische, betekenisvolle analyse. Hij merkt op dat het model de komende maanden inderdaad weer enkele keren aan de orde zal komen, onder meer in januari bij de controle van vakantieverhuur en in maart bij het convenant Woonruimteverdeling.

Bijlage: brief Bias analyse over ontwikkelde algoritmen

Tijdens de vergadering werd het toetsingskader van de bias analyse gepresenteerd. 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. Het uitvoeren van een bias analyse is dus een maatregel om te controleren of een algoritme een directe of indirecte bias bevat. De term 'bias' verwijst naar een onterecht/ onjuist verschil in inschatting/ voorspelling (en daarmee behandeling) tussen personen of groepen.

Deze methodiek is een iteratief proces dat teruggekoppeld wordt met het betreffende discriminerende algoritme dat transparant, verifieer- en controleerbaar is. Om te bepalen of een verschil gerechtvaardigd is volgt men logische redeneringen waarbij historische achtergronden, praktijkervaringen, ethische en juridische kwesties en onderzoekstechnische aspecten de revue passeren. Een bias komt men op het spoor door de uitkomsten van het model voor verschillende groepen met elkaar te vergelijken. Vervolgens speelt de vraag of het gaat om gevoelige groepen en belangrijke 'eerlijke' verschillen. Er zijn lijsten opgesteld met gevoelige classificaties van groepen en gegevens die daarover bij de gemeente bekend zijn om de bias analyse te kunnen uitvoeren.

Op het moment dat er verschillen geconstateerd worden, wordt bepaald of die aanvaardbaar zijn of niet. Daarna kan er besloten worden om het betreffende onderdeel uit het model te verwijderen of wegen te zoeken om dat wel te behouden maar de bias tegen te gaan door aanpassingen in het model of verschillende gewichten toe te kennen aan de onderdelen van de dataset. Deze beslissing moet in overleg met de inhoudelijke afdeling genomen worden. Er is een stappenplan dat gebruikt wordt om gevoelige groepen te bepalen en het formuleren van hypotheses over features die tot indirecte bias leiden, analyse van directe en indirecte bias en het wegen en mitigeren van aangetroffen bias tot het wegwerken of op een acceptabel niveau brengen van de bias.

Advies

De CPA dankt voor de goede presentatie van een interessante theoretische, betekenisvolle analyse. De commissie waardeert het dat nu enkele criteria voor gevoeligheid zijn geïdentificeerd, maar is van mening dat algoritmen daarnaast nog andere (verborgen) onderscheidingen kunnen herbergen. Daarom adviseert de CPA om samen te werken met afhankelijk van het onderwerp divers samengestelde teams van ervaringsdeskundigen om blinde vlekken op te sporen.

Ook is de CPA benieuwd er geen spanningsveld gaat optreden als men enerzijds transparantie wil bieden en anderzijds handhaving en pakkansen wil verbeteren. De commissie vraagt zich dan af of er niet een risico bestaat dat kwaadwillende daarop gaan inspelen en dat wellicht heel nieuwe vooroordelen ontstaan.

Verder wil de CPA de aandacht vestigen op de bias die het gevolg kan zijn van de mensen in het werkproces die gaandeweg hun eigen ervaring en voorkeuren inclusief mogelijke vooringenomenheid opbouwen bij de selectie van te onderzoeken situaties. De CPA meent dat het goed is om dit mee te wegen.

Tenslotte is de CPA benieuwd of er voorzien is in een audit voor dit model en adviseert de commissie nog om periodiek het model te testen met gestratificeerde steekproeven om het actueel te houden.

Verder verwijs ik u naar de notulen. Ik vertrouw er op u aldus voldoende geïnformeerd te hebben.