Algorithm Audit

Publieke standaard profileringsalgoritmes

Kwalitatieve en kwantitatieve waarborgen voor verantwoord gebruik van profileringsalgoritmes

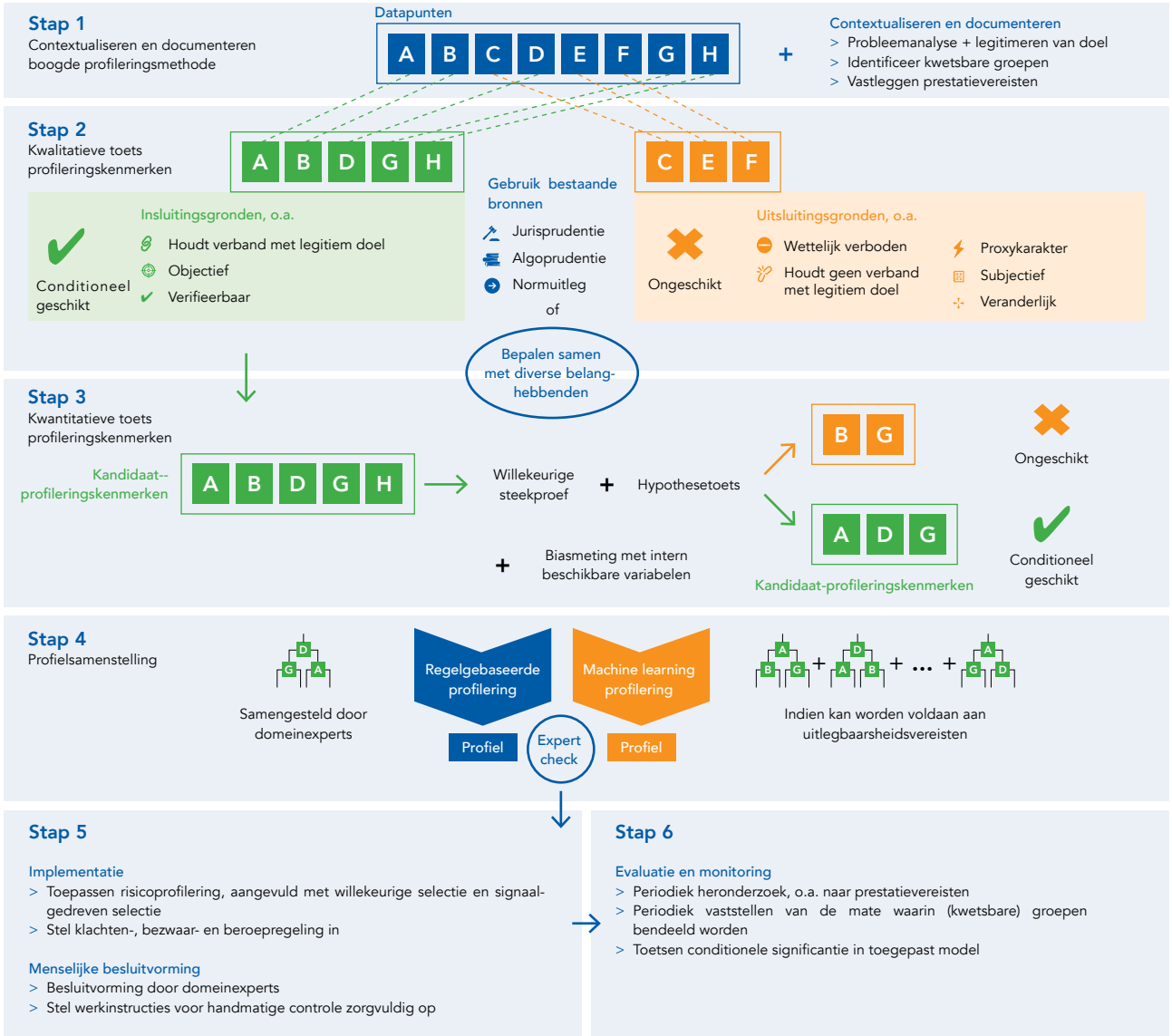
Oktober 2024

Samenvatting

Dit document doet een voorzet voor de ontwikkeling van een gestandaardiseerde werkwijze voor de verantwoorde inzet van profileringsalgoritmes in het publieke domein. Deze standaard is zowel van toepassing op zelflerende (machine learning) als regelgebaseerde algoritmes. Door algoritmes te controleren aan de hand van deze standaard kunnen (indirecte) discriminatie en andere ongewenste effecten van profilering worden bestreden. De methode volgt een stappenplan dat bestaat uit een kwalitatieve en kwantitatieve toets en aanvullende organisatorische beheersmaatregelen. De toetsingsmethoden helpen bij het invullen van open juridische normen uit het non-discriminatierecht, de Europese AI Verordening en andere wet- en regelgeving. De methoden zijn voortgekomen uit praktijkervaring met het valideren van profileringsalgoritmes. De standaard richt zich met name op toepassing in het publieke domein, maar kan ook worden gebruikt in het private domein.

Stappenplan

Onderstaand stappenplan wordt in het vervolg per stap kort toegelicht. Een volledige toelichting wordt nader uitgewerkt in de Bijsluiter – Publieke standaard profileringsalgoritmes.



Kwalitatieve en kwantitatieve waarborgen voor verantwoord gebruik van profileringsalgoritmes

Een volledige toelichting op het stappenplan wordt uitgewerkt in Bijsluiter – Publieke standaard profileringsalgoritmes.

Toelichting stappenplan

Stappenplan

Onderstaand stappenplan biedt een leidraad hoe men tot verantwoorde inzet van profileringsalgoritmes kan komen. Het biedt hiervoor echter geen garantie, omdat het afhankelijk is van de wijze waarop de stappen worden uitgevoerd en van de keuzes die hierin worden gemaakt. De vereisten uit het Algoritmekader zijn leidend voor alle verschillende fasen uit de levenscyclus van het profileringsalgoritme.[1] Deze standaard is specifiek gericht op profileringsalgoritmes. Het is daarmee een aanvulling op de Impact Assessment Mensenrechten en Algoritmes (IAMA), waarbij dit stappenplan duidelijker handvatten geeft voor het beoordelen van geschikte profileringmethoden. In deze standaard wordt niet ingegaan op vereisten voor informatiebeveiliging. Hiervoor wordt verwezen naar de Baseline Informatiebeveiliging Overheid (BIO). Voor de rechtmatige verwerking van persoonsgegevens wordt verwezen naar de Algemene Verordening Gegevensverwerking (Avg).

Stap 1 – Contextualiseren en documenteren beoogde profileringsmethode

Onderstaande stap kan eventueel worden geïntegreerd in de uitvoering van andere methodes, zoals de Impact Assessment Mensenrechten en Algoritmes (IAMA), Algoritmekader en het Onderzoekskader algoritmes van de Audit Dienst Rijk (ADR).

1.1 Beschrijf de wettelijke basis waarop handhaving en toezicht middels risicoprofilering berust.

1.2 De volgende aspecten dienen zorgvuldig te worden gemotiveerd:
- > Het probleem dat met de inzet van het algoritme moet worden opgelost;
- > De afweging of, en zo ja welk type algoritme het juiste middel is om het probleem op doelmatige wijze op te lossen.

1.3 Identificeer kwetsbare groepen in de populatie.[2] Ga na wat nadelige gevolgen voor deze groepen zijn.

1.4 Onderzoek en leg vast welke kwantitatieve maat om vooringenomenheid te meten (bias metriek) het meest relevant is voor de gegeven context.[3]

1.5 Leg prestatievereisten voor het algoritme vast.

1.6 Identificeer de beschikbare variabelen in de database die een eigenschap van natuurlijke personen of organisaties vertegenwoordigen.

---

[1] De vereisten en maatregelen uit het Algoritmekader komen overeen met eisen uit de AI Verordening wanneer het profileringsalgoritme onder de definitie van een AI-systeem valt. Wanneer de profileringsmethode niet onder de hoog-risicocategorie van de AI Verordening valt, kan het Algoritmekader alsnog worden gebruikt als beheersmaatregelenkader voor omgang met 'impactvolle algoritmes'.

[2] Beschermde gronden onder de Algemene wet gelijke behandeling (Awgb): godsdienst, levensovertuiging, politieke gezindheid, ras, geslacht, nationaliteit, hetero- of homoseksuele gerichtheid of burgerlijke staat. Maar ook gronden die formeel niet juridisch beschermd zijn, maar op basis waarvan onderscheid alsnog ethisch onwenselijk kan zijn, zoals overgewicht, opleidingsniveau en professioneel voorkomen.

[3] De relevante kwantitatieve maat verschilt per context.

Stap 2 – Kwalitatieve toets profilerings-kenmerken

2.1 Ga na of bestaande normatieve oordelen beschikbaar zijn voor juist gebruik van profileringskenmerken in een vergelijkbare context. Denk aan: jurisprudentie, algoprudentie[4] en normuitleg van toezichthouders. Indien deze informatie voorhanden is, sla stap 2.2 over en voer de daaropvolgende stappen uit aan de hand van de beschikbare oordelen.

2.2 Stel een groep van diverse belanghebbenden samen, bestaande uit onder meer de algoritme-ontwikkelaar, een domeinexpert, burgers onderworpen aan het algoritme of vertegenwoordigers daarvan, en juridische, statistische en ethische experts.[5]

2.3 Loop gezamenlijk de geïdentificeerde kenmerken in stap 1.6 bij langs en ga na of ieder kenmerk voldoet aan de volgende of andere mogelijke uitsluitingsgronden:

- > Wettelijk verboden: verboden onderscheid op basis van de Awgb[2] of onrechtmatige verwerking van persoonsgegevens in de context van het algoritme volgens de Avg;
- > Houdt geen verband met legitiem doel: kenmerk heeft geen duidelijke en inhoudelijke relatie met het nagestreefde doel;
- > Proxykarakter: kenmerk heeft sterke relatie met een kwetsbare groep[6];
- > Subjectief: kenmerk kan niet objectief worden gemeten en is gebaseerd op een subjectief waardeoordeel;
- > Veranderlijk: kenmerk is onbetrouwbaar want gebaseerd op een momentopname van een veranderlijke eigenschap.

Loop de overgebleven criteria bij langs en controleer of ze aan de volgende insluitingsgronden voldoen en waarom:

- > Houdt verband met legitiem doel: kenmerk heeft een duidelijke en inhoudelijke relatie met het nagestreefde doel;
- > Objectief: kenmerk is onafhankelijk van subjectieve waarneming of waardeoordeel;
- > Verifieerbaar: de correctheid van een kenmerk kan worden nagegaan.[7]

De eigenschappen die niet voldoen aan de uitsluitingsgronden en wel voldoen aan de insluitingsgronden gaan door naar de stap 3 en worden *kandidaat-profileringskenmerken* genoemd.

Stap 3 – Kwantitatieve toets profileringskenmerken

3.1 Trek een willekeurige steekproef uit de doelpopulatie.

3.2 Stel een hypothese op over verband tussen profileringskenmerk en het nagestreefde doel.

3.3 Pas een statistische hypothesetoets toe op de willekeurige steekproef en ga na of er een statistisch significant verband bestaat.[8]

3.4 Indien er geen sprake is van een statistisch significant verband, verwijder het kenmerk uit de verzameling kandidaat-profilerings-kenmerken.[9]

---

4 Algoprudentie: transparante collectieve oordeelsvorming over de verantwoorde inzet van algoritmes https://algorithmaudit.eu/nl/knowledge-platform/knowledge-base/white_paper_algoprudence/

5 Richtlijnen voor dit proces: https://algorithmaudit.eu/nl/algoprudence/how-we-work/#richtlijnen

6 Bijvoorbeeld: onderscheid op basis van Nederlandse taal is sterk gerelateerd aan migratieachtergrond. Onderscheid op basis van technische beroepen is sterk gerelateerd aan geslacht.

7 Spaargeld op een buitenlandse bankrekening is een voorbeeld van een kenmerk dat wel objectief, maar niet altijd verifieerbaar is. Een opgegeven bedrag is wel objectief maar kan mogelijk niet altijd geverifieerd worden door een Nederlandse overheidsinstelling.

8 Een voorbeeld kan worden gevonden in sectie 3.1 van het rapport *Vooringenomenheid voorkomen*, Algorithm Audit (2024).

9 In de bijsluiter wordt toegelicht welke statistische toets relevant is in specifieke contexten en hoe om te gaan met het toetsen van meervoudige hypotheses.

3.5 Voer, indien mogelijk, een biasmeting uit aan de hand van intern beschikbare gegevens over kwetsbare groepen.[10]

3.6 Weeg kwantitatieve inzichten voortgekomen uit de biasmeting aan de hand van kwalitatieve werkwijze uit stap 2.1-2.2.

Stap 4 – Profielsamenstelling

4.1 Stel aan de hand van overgebleven kandidaat-profileringskenmerken een risicoprofiel op. Dit profiel kan door domeinexperts of door een variabelenselectie-algoritme worden samengesteld. Documenteer en onderbouw de keuze hoe het profiel wordt samensgesteld. Overheidsorganisaties kunnen alleen variabelenselectie-algoritmen (machine learning) toepassen als kan worden voldaan aan uitlegbaarheidsvereisten.[11]

4.2 Laat een samengesteld kandidaat-risicoprofiel valideren door (externe of interne, maar wel onafhankelijke) kundige experts die niet betrokken zijn geweest bij het ontwerpproces van het risicoprofiel.

Stap 5 – Implementatie

5.1 Stel een verdeling vast in de te onderzoeken populatie tussen hoeveel personen of organisaties worden geselecteerd door het risicoprofiel, door een willekeurige steekproef, of door signaal-gedreven selectie (bijvoorbeeld gemelde klachten of andere signalen uit de organisatie). Een suggestie is een verhouding van 2:1:1.

5.2 Stel een procedure in om gehoor te geven aan de rechten van geselecteerde partijen, zoals de mogelijkheid van bezwaar, beroep en klacht. Verzeker dat klachten door de organisatie opgepikt worden.

5.3 Laat domeinexperts het uiteindelijke besluit nemen om door het risicoprofiel geselecteerde personen of organisaties daadwerkelijk te onderzoeken. Stel werkinstructies voor domeinexperts zorgvuldig op. Voorkom herhaling van profileringskenmerken in de stap van algoritmische profilering en de stap van handmatige inspectie door domeinexperts. Houd rekening met kwetsbare groepen. Documenteer waarom een persoon of organisatie uiteindelijk is geselecteerd voor een onderzoek, zodanig dat aan uitlegbaarheidsvereisten wordt voldaan.

5.4 Beheer het model met behulp van versiebeheer.

Stap 6 – Evaluatie en monitoring

6.1 Doorloop periodiek de gehele toets opnieuw met in achtneming van de eventuele gewijzigde situatie.

6.2 Stel periodiek vast welke (kwetsbare) groepen zijn geselecteerd voor controle. Gegevens over beschermde gronden kunnen hiervoor worden gebruikt, zoals intern beschikbaar binnen de organisatie of zoals opgevraagd bij het Centraal Bureau voor de Statistiek (CBS). Indien deze gegevens niet beschikbaar zijn kan een clusteringmethode worden toegepast om te onderzoeken welke groepen structureel afwijken.[12] Toets de uitkomsten aan de vastgelegde geaccepteerde hoeveelheid bias per groep.

6.3 Achterhaal de voorspellende waarde van criteria gegeven het risicoprofiel door het toetsen van conditionele significantie.[13]

6.4 Indien bij een periodieke toets niet meer aan de gestelde normen wordt voldaan, stop de profileringsmethode en archiveer het.

---

[10] Merk op dat gegevens over onder meer geslacht, taal, leeftijd, sociaaleconomische status niet altijd onder art. 9 van de Avg vallen en daardoor wel voor dit doel verwerkt kunnen worden.

[11] Uitlegbaarheidseisen voor ML-gedreven risicoprofilering kan worden gevonden in algoprudentie *Risicoprofilering heronderzoek bijstandsuitkering* (ALGO:AA:2023:02:A).

[12] Meer informatie over een biasmeting zonder toegang tot beschermde gronden: https://algorithmaudit.eu/nl/technical-tools/bdt/

[13] Wijkt het verschil in voorspelde uitkomsten significant af van nul als een profileringscriterium verwijderd wordt uit het profiel?

Over Algorithm Audit

Algorithm Audit is een Europees kennisplatform voor AI bias testing en normatieve AI-standaarden.
De doelen van de stichting zijn drieledig:

Normatieve adviescommissies	Adviseren over ethische kwesties in concrete algoritmische toepassingen door het samenbrengen van deliberatieve, diverse adviesommissies, met algoprudentie als resultaat
Technische hulpmiddelen	Implementeren en testen van technische methoden voor bias-detectie en -mitigatie, zoals onze bias detection tool
Kennisplatform	Samenbrengen van kennis en experts voor collectief leerproces over verantwoorde inzet van algoritmes, bijvoorbeeld ons AI Policy Observatory en position papers

Structurele partners van Algorithm Audit

SIDN Fonds

Het SIDN Fonds staat voor een sterk internet voor iedereen. Het Fonds investeert in projecten met lef en maatschappelijke meerwaarde, met als doel het borgen van publieke waarden online en in de digitale democratie.

European AI&Society Fund

Het European AI&Society Fund ondersteunt organisaties uit heel Europa die AI beleid vormgeven waarin mens en maatschappij centraal staan. Het fonds is een samenwerkingsverband van 14 Europese en Amerikaase filantropische organisaties.

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Het ministerie van BZK maakt zich sterk voor een democratische rechtsstaat, met een slagvaardig bestuur. Ze borgt de kernwaarden van de democratie. BZK staat voor een goed en digitaalvaardig openbaar bestuur en een overheid waar burgers op kunnen vertrouwen.

Opbouwen van *publieke kennis*

over verantwoorde AI *zonder winstoogmerk*

www.algorithmaudit.eu

www.github.com/NGO-Algorithm-Audit

info@algorithmaudit.eu

Algorithm Audit

Stichting Algorithm Audit is geregistreerd bij de
Kamer van Koophandel onder nummer 83979212