Anna van Buerenplein 1
2595 DA Den Haag
Postbus 96800
2509 JE Den Haag

www.tno.nl

T +31 88 866 00 00

TNO-rapport

TNO 2021 R10732 | 1.0.0

Technische evaluatie van het AI-algoritme dat de Gemeente Nissewaard inzet voor de opsporing van misbruik en oneigenlijk gebruik van bijstandsuitkeringen

Datum	22 juni 2021
Projectleider	Dr. ir. J.M. Tang (TNO ICT Data Science)
Aantal pagina's	44 (incl. bijlagen)
Aantal bijlagen	2
Opdrachtgever	Gemeente Nissewaard
Projectnaam	Nissewaard-casus
Projectnummer	060.47861

Anna van Buerenplein 1
2595 DA Den Haag
Postbus 96800
2509 JE Den Haag

www.tno.nl

T +31 88 866 00 00

Auteurs:
Dr.ir. J.M. Tang [1]
Dr. M.H.T. de Boer
S. Vethman MSc.

Wetenschappelijke adviseurs:
Dr. ing. C. J. Veenman
Prof. dr. T.M. van Engers
Prof. dr. ir. W. Kraaij
Prof. dr. S.A. Raaijmakers

TNO
Postbus 96800
2509 JE Den Haag

Juni 2021

[1] Corresponderende auteur. E-mailadres: jok.tang@tno.nl.

Samenvatting

Achtergrond van het onderzoek
De Gemeente Nissewaard zet een door Totta Data Lab ontwikkeld algoritme in voor risico-indicatie van misbruik of oneigenlijk gebruik van ontvangers van bijstands-uitkeringen. Het specifieke doel van dit algoritme is om de eerder gehanteerde algemene periodieke controle te vervangen, waarbij het toezicht zoveel mogelijk gericht is op de cliënten die de aandacht verdienen (ook wel genoemd: risico-gestuurd handhaven). De gemeente Nissewaard vindt het belangrijk dat het gebruik van het algoritme op correcte wijze plaatsvindt en wil de eigen werkwijze ook kritisch bekijken. Daarom heeft de Gemeente Nissewaard aan TNO gevraagd mogelijke risico's ten aanzien van het (gebruik van het) algoritme in kaart te brengen en hiertoe een technisch-inhoudelijke evaluatie van het AI-algoritme uit te voeren. Het onderzoek past in een op dit moment gevoerde brede maatschappelijke discussie over het gebruik van dit soort algoritmen. De scope van het TNO-onderzoek beperkt zich tot deze technisch-inhoudelijke evaluatie. Hierbij wordt er getoetst aan ethische richtlijnen, die als gegeven zijn gebruikt in deze evaluatie. Privacy-overwegingen, aspecten van "*profiling*" en een juridisch oordeel maken geen deel uit van het onderzoek. De werkwijze, resultaten, conclusies en aanbevelingen van de technisch-inhoudelijke evaluatie zijn beschreven in dit rapport.

Werkwijze in het onderzoek
In de technisch-inhoudelijke evaluatie zijn vier hoofdactiviteiten uitgevoerd:

- Vaststellen van de relevante evaluatieaspecten om het algoritme op te beoordelen. Dit is gedaan door het combineren van bestaande richtlijnen voor het toepassen van algoritmen van het Ministerie van Justitie en Veiligheid en die van de *high-level expert group on AI* van de Europese Commissie.
- Opstellen van een vragenlijst voor interviews, op basis van de relevante evaluatieaspecten. De interviews zijn gehouden met vertegenwoordigers van zowel de Gemeente Nissewaard als van Totta Data Lab.
- Uitvoeren van een technische analyse, waarbij gefocust is op de evaluatieaspecten "auditeerbaarheid en "doeltreffendheid en inbedding". Hierbij is gekeken naar de inputdata, de softwarecode waarin het algoritme is vastgelegd en de uitkomsten van het algoritme.
- Formuleren van de conclusies en aanbevelingen op basis van de bevindingen.

Bevindingen van het onderzoek
De positieve bevindingen ten aanzien van (de inzet van) het AI-algoritme zijn als volgt:

- In het huidige protocol van de Gemeente Nissewaard worden de risico-indicaties die door het AI-algoritme worden aangegeven gecombineerd met andere risico-indicaties, waarbij de door het AI-algoritme aangedragen te onderzoeken dossiers slecht 10-15% van de te onderzoeken populatie bedraagt. Deze risico-indicaties spelen een rol in de selectie van dossiers, maar hebben verder geen sturende invloed op het verder handmatig uitgevoerde toezicht- en handhavingsproces. De mens draagt en neemt vervolgens ook de verantwoordelijkheid voor mogelijke besluiten die uit het toezicht- en handhavingsproces voortvloeien.
- Uit interviews is gebleken dat zowel de Gemeente Nissewaard als Totta Data Lab aandacht heeft besteed aan verschillende aspecten van verantwoord gebruik van AI-algoritmen, zoals datagebruik, non-discriminatie, transparantie, verantwoording, uitlegbaarheid en publieksvoorlichting (beschreven in de richtlijnen). Naast de interviews zijn

deze aspecten niet verder onderzocht in de technische analyse van de evaluatie.

Op de volgende punten presteert het AI-algoritme onvoldoende:

- De auditeerbaarheid van het AI-algoritme is gebrekkig. Dit betekent dat het algoritme in de huidige vorm niet voldoende toetsbaar is aan inhoudelijke en procedurele eisen. De uitkomsten van het algoritme blijken namelijk niet reproduceerbaar te zijn. Dit betekent dat bij meerdere runs van dezelfde code met dezelfde instellingen er (significant) andere resultaten uitkomen. Daarnaast is de kwaliteit van de programmeercode niet op peil. Bij het implementeren van het algoritme, maar ook bij het testen van de code, zijn de *best practices* vanuit software-ontwikkeling onvoldoende toegepast. Evenmin ontbreekt het aan aanvullende documentatie voor software-ontwikkelaars en technische gebruikers van de code. Aspecten zoals reproduceerbaarheid en kwaliteit van de code zijn een voorwaarde voor het kunnen auditeren van andere relevante ethische aspecten, zoals non-discriminatie en transparantie van een algoritme-toepassing. Tijdens de interviews is geconstateerd dat er wel aandacht is besteed aan deze andere aspecten, maar deze konden niet verder worden onderzocht vanwege het gebrek aan auditeerbaarheid.
- Het is onvoldoende mogelijk om de doeltreffendheid (ook wel: effectiviteit) van het AI-algoritme vanuit een technisch oogpunt te onderzoeken. De oorzaak is dat de resultaten niet robuust zijn met betrekking tot de aanwezigheid van de willekeurige elementen in het algoritme. Dit was ook de oorzaak van het bovengenoemde probleem ten aanzien van reproduceerbaarheid van het algoritme. Uit de evaluatie blijkt dat de met behulp van het AI-algoritme vastgestelde risicoscores, inclusief de berekende volgorde van cliënten op basis van dergelijke risicoscores, niet betrouwbaar zijn. Hierdoor kan Nissewaard de uitkomsten van het algoritme niet met een gepast niveau van zekerheid interpreteren.

Hoofdconclusie van het onderzoek
Ondanks dat Nissewaard en Totta Data Lab duidelijk aandacht hebben besteed aan verschillende aspecten van verantwoord gebruik, is de hoofdconclusie van de technisch-inhoudelijke evaluatie dat het AI-algoritme in de huidige vorm van onvoldoende niveau is om verantwoord te kunnen inzetten.

Deze hoofdconclusie wijst uitsluitend op de huidige vorm van dit specifieke AI-algoritme. TNO velt nadrukkelijk geen oordeel over juridische aspecten of op principiële bezwaren tegen de inbedding van een AI-algoritme in het huidige informatiegestuurde handhavingsproces dat door de Gemeente Nissewaard wordt gehanteerd.

Aanbevelingen voor de Gemeente Nissewaard
TNO adviseert de Gemeente Nissewaard om op verschillende vlakken significante verbeteringen door te voeren in het AI-algoritme van Totta Data Lab, alvorens inzet van dit algoritme heroverwogen kan worden door de Gemeente Nissewaard. De concrete aanbevelingen zijn als volgt:

- Op methodologisch vlak dienen verbeteringen plaats te vinden, in het bijzonder op de volgende gebieden:
    - het beter omgaan met de willekeurigheid in het model, inclusief het kwantificeren van de onzekerheden hiervan in de resultaten;
    - het verbeteren van de experimentele opzet om de nauwkeurigheid van het algoritme te meten.
- Op code-technisch vlak dienen de testmethodiek, de softwareconfiguratie en de documentatie verbeterd te worden.
- Op procesmatig vlak dienen de adequaatheid van het AI-algoritme als risicovoorspeller en daarvan afgeleid de doeltreffendheid van het AI-

algoritme te worden vastgesteld en structureel te worden gemonitord. Hierbij kan er gebruik worden gemaakt van een ingelaste observatieperiode. In die periode kan de adequaatheid van de werking van een verbeterd AI-algoritme verder worden onderzocht. De Gemeente Nissewaard wordt aangeraden om niet eerder dan aan het eind van deze periode de inzet van het AI-algoritme, als onderdeel van het informatiegestuurde handhavingsproces, te heroverwegen.

Inhoudsopgave

Samenvatting 3

1 Inleiding 7
1.1 Risicogestuurd handhaven ondersteund door AI in Nissewaard 7
1.2 Randvoorwaarden voor AI-toepassing in risicogestuurd handhaven 8
1.3 Projectafbakening 9
1.4 Leeswijzer 10

2 Risicogestuurde handhaving 11
2.1 Traditioneel risicogestuurde handhaving binnen Nissewaard 11
2.2 Huidige werkwijze met AI 12
2.3 Aandachtspunten bij het gebruik van AI-algoritmen 13

3 Aanpak van de technisch-inhoudelijk evaluatie 15
3.1 Combineren van richtlijnen en opstellen van vragenlijst 15
3.2 Interviews 16
3.3 Technische analyse van data, code en uitkomsten 17

4 Resultaten van de technisch-inhoudelijke evaluatie 19
4.1 Resultaten van het combineren van richtlijnen 19
4.2 Resultaten vanuit de interviews en technische analyse 20

5 Conclusies en aanbevelingen 27
5.1 Hoofdbevindingen 27
5.2 Hoofdconclusie 28
5.3 Aanbevelingen 28

6 Referenties 30

Bijlage 1: Vragenlijst voor interviews t.a.v. de richtlijnen voor betrouwbare AI 32

Bijlage 2: Resultaten van de interviews per richtlijn 42

1 Inleiding

In dit rapport presenteren we de uitkomsten van de technische evaluatie van het AI-algoritme dat de Gemeente Nissewaard (hierna: Nissewaard) gebruikt voor de opsporing van misbruik en oneigenlijk gebruik van bijstandsuitkeringen. Hierbij staat AI voor Artificial Intelligence, oftewel kunstmatige intelligentie.

In dit hoofdstuk beschrijven we achtereenvolgens de setting van risicogestuurd handhaven in Nissewaard (Sectie 1.1), de randvoorwaarden voor de toepassing van AI in de opsporing van misbruik en oneigenlijk gebruik (Sectie 1.2), de projectafbakening (Sectie 1.3) en de leeswijzer (Sectie 1.4).

1.1 Risicogestuurd handhaven ondersteund door AI in Nissewaard

Gemeenten in Nederland zijn uitvoerders van de Participatiewet. Zij ondersteunen bij arbeidsinschakeling en verstrekken bijstand. Gemeenten zijn daarmee verantwoordelijk voor de beoordeling of burgers recht hebben op een uitkering. Onderdeel van deze taak is ook het toezien op de rechtmatigheid van een verstrekte uitkering. Gemeenten hebben vergaande rechten om informatie op te vragen bij instanties. Burgers hebben een informatieplicht om relevante wijzigingen in hun situatie door te geven en op informatieverzoeken te reageren. Nalatigheid of oneigenlijk gebruik kan leiden tot stopzetting van de uitkering en het opleggen van een terugvordering en/of boete. Hierbij dient opgemerkt te worden dat er niet in alle gevallen van oneigenlijk gebruik sprake is van fraude, maar dat soms ook onbedoeld sprake kan zijn van onrechtmatig gebruik wat bij niet tijdig signaleren tot grote individuele problemen kan leiden op de langere termijn. Hierbij is het ook waardevol voor bijstandsgerechtigden die onbedoeld te veel inkomsten van de gemeente ontvangen om daar tijdig achter te komen. Dit voorkomt een hogere terugvordering. De toekenning en hoogte van een uitkering zijn sterk afhankelijk van de situatie van de bijstandsgerechtigde. Een wijziging daarin kan leiden tot een aanpassing of intrekking, b.v. bij de geboorte van een kind of inkomsten uit arbeid. Als een burger verzuimt dergelijke aanpassingen door te geven, kan er mogelijk sprake zijn van misbruik of oneigenlijk gebruik.

Om te voorkomen dat de ontvangers van bijstandsuitkeringen veelvuldig moeten worden gecontroleerd – wat kostbaar is en ook een administratieve last vormt voor de uitkeringsgerechtigden – heeft Nissewaard gekozen voor risicogestuurde handhaving (ook wel informatiegestuurde handhaving genoemd). Risicogestuurde handhaving is een vorm van handhaving waarin het zwaartepunt van de capaciteit wordt ingezet op die plekken waar de risico's, op oneigenlijk gebruik, het grootst zijn. Risicogestuurd werken vindt ook bij verschillende andere in Nederland plaats, b.v. bij GGD's, politie en inspecties, zie ook [15]. Naast risicogestuurde handhaving hanteert Nissewaard ook themacontroles (zie ook Sectie 2.1).

De risicogestuurde handhaving binnen Nissewaard kent de volgende drie stappen die door ambtenaren wordt uitgevoerd.

- In de eerste stap worden dossiers van cliënten verzameld met een themacontrole of een verhoogde risico-indicatie, bijvoorbeeld door middel van een melding of een automatische risico-inschatting door een algoritme.
- In de tweede stap worden deze dossiers door sociaal rechercheurs nader beoordeeld op onderzoekwaardigheid.
- In de derde stap worden de onderzoekwaardige dossiers ingepland op basis van de beschikbare onderzoekscapaciteit (zie Sectie 2.1 voor meer detail).

Bij Nissewaard gebeurt het verzamelen van de dossiers (stap 1) niet alleen op basis van meldingen of themacontroles, maar ook met het AI-algoritme dat is ontwikkeld door Totta Data Lab B.V. (hierna: TDL). Met een AI-algoritme wordt bedoeld een rekenmodel oftewel een geautomatiseerde analyse van gegevens, waarbij AI-technologie wordt ingezet.

Het doel van het AI-algoritme is om op basis van een aantal data-elementen te leren bij welke bijstandsontvangers er een verhoogd risico op misbruik of oneigenlijk gebruik van bijstand bestaat. Het AI-algoritme probeert daartoe patronen in een aantal data-elementen van bijstandsontvangers te vinden die wijzen op een verhoogd risico op misbruik of oneigenlijk gebruik van bijstand. Het algoritme bepaalt op basis van deze patronen een score voor elke cliënt die mogelijk correspondeert met onrechtmatig gedrag. Deze scores worden vervolgens gebruikt om de cliënten te ordenen. De hoogst scorende cliënten worden meegenomen in de selectie van de door ambtenaren nader te beoordelen gevallen (stap 2). In de technische documentatie van TDL [1] is uitgelegd welke data gebruikt wordt en hoe dit model tot stand is gekomen.

Het AI-algoritme wordt verder niet gebruikt voor het sturen van de wijze waarop de controles worden verricht (stap 2 en 3).

1.2 Randvoorwaarden voor AI-toepassing in risicogestuurd handhaven

Het gebruik van AI-algoritmen, in het bijzonder in bestuursrechtelijke context, ligt recentelijk onder een vergrootglas, zie onder meer het ongevraagd advies van de Raad van State [6], en recente gerechtelijke uitspraken in relatie tot b.v. SyRI [8] en de e-Screener [14]. Zowel in de wetenschap als in andere delen van de maatschappij wordt momenteel een debat gevoerd over het gebruik van dergelijke algoritmen. Veelgenoemde punten van kritiek in die discussie hebben betrekking op niet-proportioneel datagebruik en het door de inzet van algoritmen opgewekte gevoel dat burgers bij voorbaat verdacht zijn (zie ook Sectie 2.3 en [16]). Ook Nissewaard heeft deze ‘maatschappelijke druk’ ervaren. Onder andere NRC, AD, Binnenlands Bestuur en NOS hebben sinds 2018 gepubliceerd over gebruik van het AI-algoritme door Nissewaard [3, 4, 11, 19]. In deze artikelen zijn door partijen zoals FNV en Platform Bescherming Burgerrechten vraagtekens gezet bij het voldoende controleerbaar en inzichtelijk zijn van het algoritme. Ook wordt geclaimd dat verschillende databronnen oneigenlijk gebruikt zouden zijn en zou het AI-algoritme een te grote inbreuk maken op de privacy van burgers (zie ook [3, 4, 5]). Tenslotte worden er parallellen getrokken tussen het gebruik van AI-algoritmen door Nissewaard en de toeslagenaffaire [12].

In het publieke debat rond AI-ondersteund toezicht zijn verschillende invalshoeken te onderscheiden: a) ethische discussies (b.v. individueel belang versus maatschappelijk belang); b) juridische vragen: zijn de methoden legitiem; c) technische/methodologische vragen: in hoeverre is de methode reproduceerbaar en uitlegbaar.

De ethische, juridische en technische discussie over verantwoorde toepassing van AI wordt al enkele jaren gevoerd en heeft geleid tot een aantal richtlijnen op nationaal en Europees niveau, zoals de richtlijnen van het Ministerie van Justitie en Veiligheid (hierna: MinJenV) en die van de *high-level expert group on AI* van de Europese Commissie (hierna: EC), zie [7, 9, 18]. [2]

Nissewaard vindt het belangrijk dat het gebruik van het algoritme op correcte wijze plaatsvindt en wil de eigen werkwijze ook kritisch bekijken. Het in kaart brengen van mogelijke risico's ten aanzien van het (gebruik van het) algoritme is hierbij van belang. Dit is ook mede aangewakkerd door een maatschappelijke discussie over het gebruik van dit soort algoritmen. Daarom heeft de Gemeente Nissewaard aan TNO als onafhankelijke onderzoeksinstelling gevraagd om een technisch-inhoudelijke evaluatie van het AI-algoritme uit te voeren. In deze evaluatie wordt in kaart gebracht welke technische aspecten een risico vormen voor verantwoord gebruik van het bovengenoemde AI-algoritme als onderdeel van risicogestuurd handhaven. De scope van het onderzoek richt zich op de evaluatie, waarbij een technisch oordeel over (de toepassing van) het gehanteerde algoritme wordt gegeven. TNO zal verder geen juridisch oordeel uitspreken over de rechtmatige inzet van het AI-algoritme. Dit onderzoek wordt in dit rapport de "*Nissewaard-casus*" genoemd.

1.3 Projectafbakening

Het doel van de Nissewaard-casus is om een technisch-inhoudelijke evaluatie uit te voeren op (de code-implementatie van) het AI-algoritme van TDL langs thans bestaande richtlijnen rondom betrouwbare algoritmen. Merk overigens op dat deze richtlijnen recent opgesteld zijn, namelijk in 20192019, terwijl het algoritme al voor die tijd door TDL is ontwikkeld en door Nissewaard in gebruik is genomen is.

Ten behoeve van deze evaluatie zijn interviews met verschillende betrokkenen gehouden en is het AI-algoritme en de wijze waarop het ontwikkeld, getest en ingezet is aan een technische analyse onderworpen. Bij deze technische analyse is er specifiek stilgestaan bij de volgende twee evaluatieaspecten:

- *Auditeerbaarheid*: De ontwikkeling en het gebruik van de AI-toepassing heeft zorgvuldige onderbouwing van de daarbij gehanteerde uitgangspunten en gemaakte ontwerpkeuzen nodig en vergt gedetailleerde documentatie. Het dient mogelijk te zijn om te controleren of het algoritme voldoet aan inhoudelijke en procedurele eisen. Dit is essentieel voor transparantie en het kunnen afleggen van verantwoording over het toepassen van het algoritme.
- *Doeltreffendheid en inbedding*: Het gebruik van een algoritme dient met enige zekerheid bij te dragen aan de doeltreffendheid van het proces waarin het is ingebed (ook wel: effectief). In geval van Nissewaard betreft deze doeltreffendheid dat van de door het algoritme aangeduide risicovolle cliënten, na menselijke controle, daadwerkelijk kan worden vastgesteld dat sprake is van oneigenlijk gebruik of misbruik. Doeltreffendheid kan worden gemeten door de nauwkeurigheid van het algoritme in de testprocedure van de code, aangevuld door validatie van het proces waarin algoritme is ingebed. Voor deze validatie zijn derhalve de resultaten van menselijke controle op de geïndiceerde risicovolle dossiers nodig.

[2] Verder is recentelijk een voorstel verschenen voor nieuwe EU-wetgeving t.a.v. AI, zie [17]. Het algoritme dat door Nissewaard wordt ingezet in het informatiegestuurd toezicht moet waarschijnlijk worden aangemerkt als 'High-Risk AI' volgens punt 5 in Annex III (p.4) van [17]. Het verdient daarom aanbeveling te toetsen of aan alle eisen die aan dergelijke AI-toepassingen worden gesteld wordt voldaan. Dit valt buiten de scope van de technisch-inhoudelijke evaluatie zoals in dit rapport beschreven.

De evaluatieaspecten auditeerbaarheid en doeltreffendheid zijn een voorwaarde om andere relevante ethische aspecten, zoals bias en transparantie, van een algoritme-toepassing te evalueren. In de Nissewaard-casus is ook als input meegenomen de rechterlijke uitspraak t.a.v. Systeem Risico Indicatie oftewel SyRI (zie [8]) en het rapport t.a.v. de toeslagenaffaire (zie [12]), die aansluiten op de hierboven beschreven projectafbakening. Merk op dat verder in de Nissewaard-casus geen juridisch oordeel wordt geveld over noch deze input noch de connectie met de Nissewaard-casus.

Al voor aanvang van het onderzoek de Nissewaard-casus is door juridische experts van Nissewaard en Stimulansz een aantal ethische, juridische, transparantie- en privacy-gerelateerde randvoorwaarden opgesteld. Bij het formuleren van deze randvoorwaarden is onder meer gekeken naar de van toepassing zijnde sociale-zekerheidswetgeving en de AVG. Deze randvoorwaarden vormden het kader waarbinnen het AI-algoritme is ontwikkeld en wordt toegepast. Voor het onderzoek heeft TNO zich voornamelijk gericht op eisen zoals geformuleerd in recent gepubliceerde richtlijnen, te weten die van de EC en die van het MinJenV (inclusief de herziene versie verschenen in maart 2021), zie ook [7, 9, 18].
Deze geformuleerde randvoorwaarden betreffen de volgende twee punten:

- TNO voert de evaluatie uit aan de hand van materiaal dat door Nissewaard en TDL beschikbaar is gesteld aan TNO. Dit materiaal betreft vereiste data, code, documentatie, testen e.d. die nodig zijn voor de technische evaluatie en voorafgaand aan de casus met TNO is gedeeld. Dit materiaal met gegevens staat vast voor de rest van de casus; de technische evaluatie is gebaseerd op deze gegevens, waarbij is aangenomen dat deze gegeven juist zijn.
- TNO voert de evaluatie uit waarbij gefocust wordt op risico's die voortvloeien uit de gehanteerde AI-technologie. De legitimiteit van het gebruik van de data als input voor het AI-algoritme voor het doel waartoe Nissewaard dit gebruikt valt buiten de scope van dit onderzoek.

TNO kijkt in haar onderzoeksprogramma onder andere naar innovaties gerelateerd aan betrouwbare AI voor publieke instanties. De Nissewaard-casus sluit hier goed bij aan.

1.4 Leeswijzer

Hoofdstuk 2 beschrijft de context waarbinnen de huidige AI-toepassing wordt ontwikkeld en ingezet. In hoofdstuk 3 beschrijven we de opzet van de technisch-inhoudelijke evaluatie van het gebruikte AI-algoritme. Hoofdstuk 4 bevat de resultaten van deze evaluatie. In hoofdstuk 5 presenteren we de conclusies en aanbevelingen van de Nissewaard-casus.

2 Risicogestuurde handhaving

In dit hoofdstuk schetsen we de context van de wijze waarop traditioneel het toezicht bij Nissewaard was ingericht, beschrijven we het AI-algoritme dat momenteel gebruikt wordt door Nissewaard en geven we een overzicht van aandachtspunten bij de ontwikkeling en inzet van dit soort algoritmen. De aanpak en de resultaten van de technisch-inhoudelijke evaluatie zijn te vinden in Hoofdstuk 3 en 4, respectievelijk.

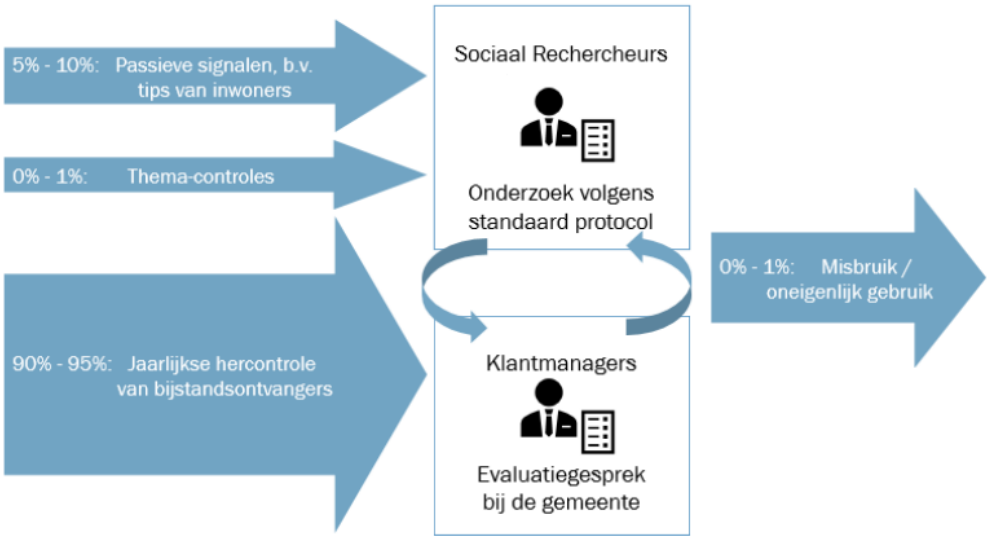
2.1 Traditioneel risicogestuurde handhaving binnen Nissewaard

Voordat informatiegestuurd toezicht werd ingevoerd binnen Nissewaard, vond handhaving op twee verschillende manieren plaats [13]:

1. Controle op basis van signalen, ook wel meldingen genoemd. Deze signalen zijn tips over mogelijke misbruik of oneigenlijk gebruik, vanuit bijvoorbeeld inwoners, het OM, het UWV en de SVB. Ook in het huidige proces van handhaving maakt het acteren op basis van deze signalen het merendeel (80-90%) van het werk voor sociaal rechercheurs uit.
2. Periodieke controles. In het traditioneel handhaven van de bijstandswetten werden alle burgers elk jaar of elke paar jaren verzocht documenten in te vullen en zich te melden bij het gemeentehuis voor een controle.

Bij risicogestuurde handhaving is daar een manier aan toegevoegd:

3. Thema-controles. In thema-controles wordt er gekeken naar een bepaald thema, zoals gedeeltelijke inkomsten. In dit voorbeeld worden mogelijke gedeeltelijke zwarte inkomsten uit werkzaamheden die niet zijn gemeld in de bijstand onderzocht.

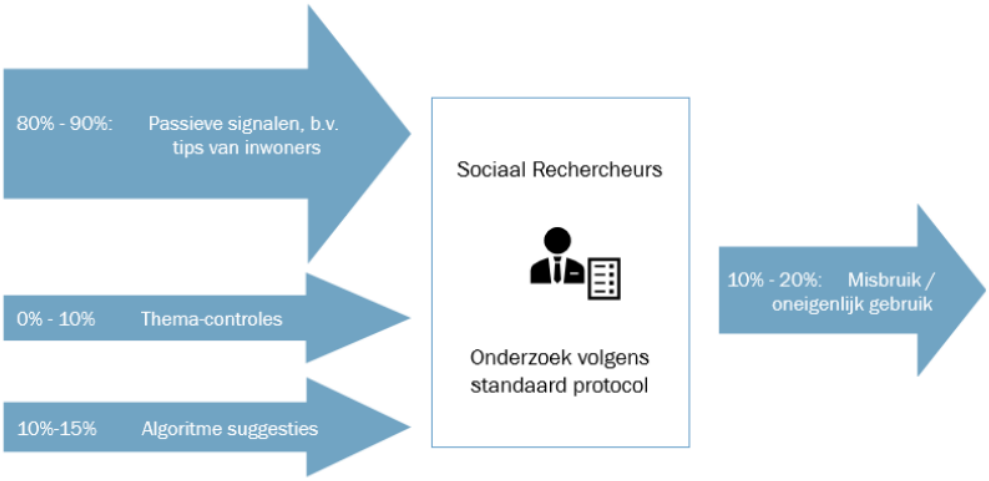


Figuur 1: De aanpak van risicogestuurd toezicht van Nissewaard voordat het AI-algoritme is ingezet. De vermelde percentages zijn schattingen.

Sociaal rechercheurs en klantmanagers, beide ambtenaren met een toezicht- en handhavingstaak, spelen een cruciale rol bij het tijdig constateren en tegengaan van misbruik en oneigenlijk gebruik van de bijstand, zie Figuur 1. De pijlen aan de linkerkant van het figuur geven aan dat sociaal rechercheurs vooral ingezet zijn op passieve signalen en thema-controles en klantmanagers vooral op de jaarlijkse hercontrole. De pijlen tussen de ambtenaren in het figuur geven aan dat deze ambtenaren gezamenlijk verantwoordelijk zijn voor het signaleren en optreden

tegen misbruik en oneigenlijk gebruik en elkaar daarbij waar mogelijk ondersteunen.

In de huidige werkwijze heeft risicogericht werken de plaats van deze periodieke controles overgenomen. Hierin wordt een AI-algoritme gebruikt waarmee risico-indicaties worden bepaald, waarna de meest risicovolle dossiers nader worden onderzocht. In de huidige opzet van het risicogestuurde toezicht worden dus, naast de cliënten die gecontroleerd worden op basis van signalen of omdat zij in de thema-controles vallen, alleen cliënten waarvoor een verhoogd risico is vastgesteld nader onderzocht. De voor veel uitkeringsgerechtigden belastende periodieke controles zijn hiermee komen te vervallen. Deze methodiek is samengevat in Figuur 2, zie de volgende sectie voor meer details.



Figuur 2: De huidige aanpak van risicogestuurd toezicht van Nissewaard inclusief de toepassing van het AI-algoritme. De vermelde percentages zijn schattingen.

2.2 Huidige werkwijze met AI

In de huidige werkwijze wordt een AI-algoritme ingezet, waarmee gepoogd wordt patronen in de gegevens van ontvangers van bijstand te ontdekken die duiden op een verhoogd risico op misbruik of oneigenlijk gebruik van bijstand. Op basis van deze patronen kent het model voor elke cliënt een risicoscore tussen 0 (laag risico) en 1 (hoog risico) toe. Aan de hand van deze score worden de cliënten gerangschikt van hoog naar laag risico, waarbij de meeste risicovolle cliënten mogelijk onderzocht worden door ambtenaren. Zowel het ontwikkelen van het AI-algoritme als het toepassen daarvan op een periodiek door Nissewaard geleverd databestand gebeurt door TDL. Vervolgens deelt TDL de gevonden ranking met Nissewaard, waarna Nissewaard de cliënten met de hoogste risicoscores meeneemt in de verzameling nader te onderzoeken cliënten.

Data

De databronnen waarop het AI-algoritme getraind wordt zijn een selectie uit de data die Nissewaard ten behoeve van haar toezicht en handhavingsprocessen regulier verzamelt. De selectie van deze databronnen is, bij aanvang van de ontwikkeling van het AI-algoritme, in overleg tussen Nissewaard en TDL bepaald. Meer informatie over de data en de selectiekeuzes is te vinden in de technische documentatie van TDL [1] en de infographic van Nissewaard [13]. TNO heeft voor haar evaluatie-onderzoek gepseudonimiseerde data ontvangen van Q4 2020, met specifieke datum 22 december 2020. Deze data betrof dezelfde selectie van data die TDL van Nissewaard heeft ontvangen om het AI-algoritme mee te draaien.

AI-algoritme
Het AI-algoritme heeft als doel om van de huidige populatie van bijstands-ontvangers het risico op misbruik en oneigenlijk gebruik van bijstand vast te stellen en vervolgens de cliënten op basis van deze risico's te rangschikken. TDL gebruikt daarvoor een zogenaamd "*rule-fit-random-forest-ensemble*" AI-algoritme [16]. Gedetailleerde informatie hierover is te vinden in de technische documentatie [1]. Voor de technisch-inhoudelijke evaluatie kon TNO over de programmeercode beschikken waarin het AI-model door TDL is geïmplementeerd. Het door TDL ontwikkelde algoritme werkt volgens een '*supervised learning*' methode, waarbij voor de training van deze methode een dataset wordt gebruikt met bijstandsontvangers waar eerder misbruik of oneigenlijk gebruik is vastgesteld (ook wel genoemd: 'de positieven') en gegevens van alle andere bijstandsontvangers (deze worden behandeld als 'de overigen'). Na toepassing van het getrainde algoritme op data van alle bijstandsontvangers wordt een lijst met de bijstandsontvangers met de hoogste risicoscore na menselijke controle aan de sociaal rechercheurs overhandigd om de standaardprocedure uit te voeren (zie Sectie 1.1). Deze menselijke controle wordt uitgevoerd in samenwerking tussen TDL en Nissewaard. Daarbij wordt onder meer beoordeeld of een specifieke demografische groep (b.v. vrouwen tegenover mannen) niet te vaak de hoogste score ontvangt en of de uitworp niet reeds eerder onderzochte bijstandontvangers bevat. De ordening heeft alleen invloed op welke bijstandsontvangers verder worden onderzocht. De specifieke risicoscore heeft geen invloed op de wijze waarop het nader onderzoek wordt uitgevoerd. De risicoscore bepaalt alleen de relatieve positie van een cliënt in de ordening. Het aantal cliënten dat nader zal worden onderzocht wordt bepaald door de beschikbare capaciteit van de sociaal rechercheurs.

Maatstaven in gebruik voor het AI-algoritme
De beoordeling van het AI-algoritme gebeurt door middel van: 1) een kwantitatieve maat voor nauwkeurigheid en 2) een menselijke controle op de suggesties van het algoritme. De kwantitatieve check bestaat uit het gebruik van een "training en test split", waarbij de beschikbare data in tweeën wordt gedeeld. Door het AI-algoritme patronen te laten leren op het ene deel (traininggedeelte), kan men daarna op het andere deel (testgedeelte) testen of met de gevonden patronen met voldoende nauwkeurigheid gevallen kunnen worden geïdentificeerd waarvan eerder is vastgesteld dat sprake was van misbruik of oneigenlijk gebruik. In de Nissewaard-casus wordt de nauwkeurigheid vooral beoordeeld door te kijken hoeveel gevallen van misbruik of oneigenlijk gebruik in de top-10 en top-100 van het testgedeelte staan.

2.3 Aandachtspunten bij het gebruik van AI-algoritmen

Als input voor de technisch-inhoudelijke evaluatie is er een korte analyse gedaan van de zogenaamde 'aandachtspunten' die gelden bij het gebruik van een AI-algoritme in een proces als risicogerichte handhaving. Hierbij is er gekeken naar verschillende bronnen zoals nieuwsartikelen. Zoals aangegeven in Sectie 1.2, wordt er zowel in de wetenschap als in andere delen van de maatschappij een debat gevoerd over het gebruik van dergelijke algoritmen. We merken hierbij op dat deze aandachtspunten aan verandering onderhevig is, vanwege de actualiteit hiervan in de wetenschap en maatschappij.

De meeste aandachtspunten zijn direct gerelateerd aan de uitlegbaarheid en doeltreffendheid alsmede de publieksvoorlichting. Aandachtspunten bevatten o.a. zorgen over het gebruik van persoonsgegevens, het trainen op eerder bekende gevallen, het aantal opgespoorde fraudeurs met de nieuwe methodiek, de begrijpelijkheid van het algoritme voor de burger en de ondoorzichtigheid en controleerbaarheid van opsporingssystemen zoals SyRI [8]. Hierbij wordt in de

media dus expliciet de link gemaakt met SyRI en de uitgebreide bespreking van dát systeem tijdens de rechtszaak over de legitimiteit van dat systeem [8, 21].

SyRI werd, tot het gebruik ervan werd aangevochten, door de overheid gebruikt voor de bestrijding van fraude op het terrein van uitkeringen, toeslagen en belastingen. De in die rechtszaak genoemde technische aantijgingen richtten zich op de kwaliteit en accuraatheid van het algoritme, disproportioneel datagebruik, strijdigheid met het beginsel van dataminimalisatie zoals benoemd in de AVG**,** gebrek aan informatie over het gebruik van risicoprofielen en een risicomodel, het onvoldoende inzichtelijk en controleerbaar zijn van SyRI en risico op uitsluiting en discriminatie. Een aantal van deze argumenten is ook te vinden in de door de media genoemde aandachtspunten voor Nissewaard. In vergelijking tussen SyRI en het AI-algoritme van TDL, gebruikt Nissewaard alleen de data die ten behoeve van het beoordelen van de rechtmatigheid van bijstandsuitkeringen noodzakelijk is en die reeds ten behoeve van dat proces wordt verzameld en verwerkt. Bovendien is de inzet van het algoritme in geval van Nissewaard beperkt tot het beoordelen van risico's van misbruik en oneigenlijk gebruik van verstrekte bijstandsuitkeringen, en worden daarover geen automatische besluiten genomen. Tenslotte is er technische documentatie aanwezig en heeft zowel Nissewaard als TDL daarin expliciet rekening gehouden met de uitlegbaarheid van de resultaten van het AI-algoritme.

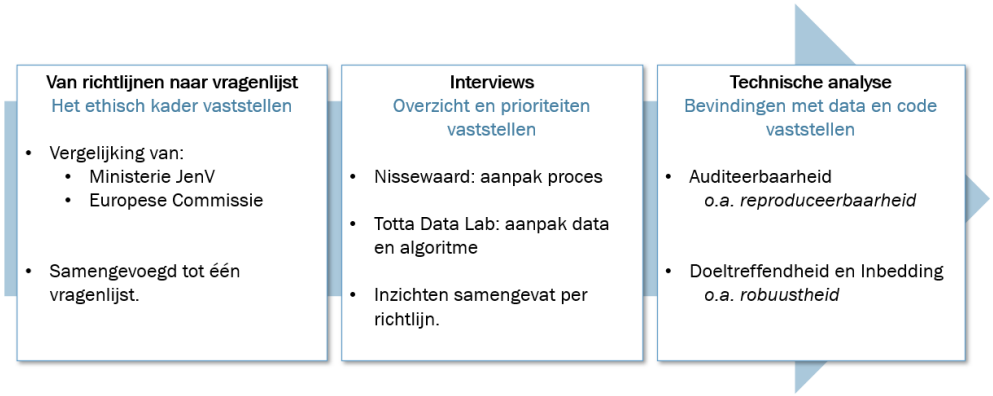
Ook uit de toeslagenaffaire zijn lessen te trekken [12]. De belangrijkste aandachtspunten die in de toeslagenaffaire worden genoemd zijn overigens organisatorisch van aard. Een probleem van het toeslagensysteem was dat mensen die door het daarbij gebruikte algoritme werden geoormerkt als potentiële fraudeur, daar ook niet meer vanaf kwamen, ook als dit aantoonbaar niet het geval was. Dit is in strijd met de AVG, waarin staat dat men het recht heeft om correcties aan te (laten) brengen als de gegevens onjuist zijn. Een technisch aandachtspunt in de toeslagenaffaire, en daarin schuilt enige overeenkomst met de Nissewaard-casus, is dat er gebruik gemaakt is van een risicoclassificatiemodel. Het algoritme dat voor het toeslagensysteem werd ingezet gebruikte daarvoor indicatoren, die vanuit het voorkomen van bias ongewenst zijn, zoals (dubbele) nationaliteit. De indicator "nationaliteit" wordt niet gebruikt in de dataset voor het trainen en toepassen van het AI-algoritme in de Nissewaard-casus. De indicatoren die hier wel zijn gebruikt staan vermeld in [1].

Door de tijdens het onderzoek naar voren komende problemen met betrekking tot de robuustheid van het algoritme (zie Hoofdstuk 4 voor meer details), heeft er geen verder technisch onderzoek plaatsgevonden in de Nissewaard-casus omtrent het risico op uitsluiting en discriminatie.

3 Aanpak van de technisch-inhoudelijk evaluatie

De technisch-inhoudelijke evaluatie is uitgevoerd door TNO-experts, met ondersteuning door de partijen Nissewaard en TDL. Om de kwaliteit van het onderzoek te kunnen waarborgen is het met het onderzoek belaste TNO-projectteam bijgestaan door een wetenschappelijke klankbordgroep (samengesteld uit experts van TNO en CBS op het gebied van betrouwbare AI) en een gemeentelijke klankbordgroep (met daarin VNG en vertegenwoordigers van enkele andere gemeenten naast Nissewaard). Ook is met het Ministerie van Justitie en Veiligheid (MinJenV) het toepassen van de door hen opgestelde richtlijnen t.a.v. betrouwbare algoritmen besproken.

De aanpak van de technisch-inhoudelijke evaluatie bestaat uit drie stappen; zie Figuur 3 voor een schets van deze drie stappen. In het onderzoek is een vragenlijst ontwikkeld gebaseerd op de richtlijnen van het MinJenV en de EC (Sectie 3.1). Deze vragenlijst is gebruikt om de betrokken partijen TDL en Nissewaard te interviewen (Sectie 3.2). Vervolgens is een technische analyse van de data en code gedaan (Sectie 3.3) op basis van twee gekozen evaluatieaspecten “auditeerbaarheid” en “doeltreffendheid en inbedding”. Deze evaluatieaspecten zijn gekozen aan de hand van de interviewresultaten en advies van de bovengenoemde klankbordgroepen.



Figuur 3: Overzicht van de driedelige aanpak van de technisch-inhoudelijke evaluatie.

3.1 Combineren van richtlijnen en opstellen van vragenlijst

Het door TNO gehanteerde ethische kader is, zoals eerder aangegeven, gebaseerd op de recent gepubliceerde richtlijnen van de EC en van MinJenV (inclusief de herziene versie verschenen in maart 2021), zie ook [7, 9, 18]. Dit kader heeft TNO in samenspraak met Nissewaard vastgesteld. Deze richtlijnen geven handvatten voor de ontwikkeling en het gebruik van algoritmen door de overheid en ten behoeve van de publieksvoorlichting. Om de exacte scope van de technische aspecten van de evaluatie te bepalen zijn de richtlijnen vergeleken in de context van de Nissewaard-casus. De vergelijking is uitgevoerd op hoofdlijnen omtrent de doelgroep, focus en toepasbaarheid en op detailniveau aan de hand van de onderwerpen in de vragenlijsten van de richtlijnen. Vervolgens hebben we de keuze gemaakt in samenwerking met de belanghebbenden en de klankbordgroepen om ze deels te combineren tot het ethische kader, dat de basis vormt voor de vragenlijst voor de interviews en de verdere technische evaluatie. We komen hierop terug in Sectie 4.1.

3.2 Interviews

Op basis van de opgestelde vragenlijst zijn interviews gehouden met vertegenwoordigers van zowel TDL als Nissewaard. In Sectie 4.2 gaan we verder in op de uitkomsten van de interviews

Zoals hierboven genoemd is op basis van o.a. de interview-uitkomsten gekozen om de technische evaluatie te richten op de evaluatieaspecten “auditeerbaarheid” en “doeltreffendheid en inbedding”. Auditeerbaarheid betreft de mate waarin het mogelijk is om het algoritme te toetsen aan inhoudelijke en procedurele eisen. Dit aspect is relevant, omdat het een voorwaarde is om de andere evaluatieaspecten zoals uitlegbaarheid te kunnen beoordelen. Doeltreffendheid en inbedding betreffen de mate waarin het algoritme correct werkt en de wijze waarop het operationeel wordt ingezet. Deze aspecten zijn relevant om de bruikbaarheid van het algoritme te kunnen vaststellen. Pas als de bruikbaarheid van voldoende niveau is, zijn de overige evaluatieaspecten relevant om verder te onderzoeken.

De evaluatieaspecten en de referentie naar de specifieke vragen of statements vanuit de vragenlijst zijn te vinden in Tabel 1. Meer toelichting over de evaluatieaspecten en de gehanteerde analyse-aanpak zijn beschreven in Sectie 3.3.

Evaluatieaspect	**Specifieke vragen vanuit de Richtlijnen**
Auditeerbaarheid: Herhaalbaarheid	• Kan men de code van het model hergebruiken en precies nadoen hoe de eerste modelresultaten waren, mits de data uit die tijd kunnen en mogen worden gebruikt?
Auditeerbaarheid: Reproduceerbaarheid	• Zijn de resultaten van het algoritme reproduceerbaar?
Auditeerbaarheid: Versiebeheer	• Is er gezorgd voor versiebeheer op de code? • Zijn alle wijzigingen aan de code bijgehouden, waardoor het mogelijk is om terug te gaan naar de eerste versie van het model?
Auditeerbaarheid: Technische documentatie	• Is de gebruikte data en de relevantie ervan gedocumenteerd? • Tevens is hier van belang het documenteren/annoteren in de code wat een routine doet of moet doen om een goede code-review uit te kunnen laten voeren door b.v. een andere analist: o Zijn alle stappen in de analyse gecodeerd? o Is ervoor gezorgd dat er geen handmatige acties nodig zijn?
Doeltreffendheid en inbedding: Robuustheid	• Verandering grootte van en tijdsinvloed op data en de invloed op de uitkomsten. o Wordt het algoritme getest wanneer er sprake is van 1 of meerdere van de onderstaande punten? 1. Verandering van data 2. Afname van de data 3. Toename van de data 4. Verandering van de dataverdeling
Doeltreffendheid en inbedding: Testprocedures	• Wordt het algoritme getest op basis van testcases/scenario’s? • Zijn alle scripts, waar mogelijk, afzonderlijk van elkaar getest? Zijn alle scripts in combinatie met elkaar getest? Is dit een standaard onderdeel van de modelontwikkeling? • Kunnen de modules op correcte functionaliteit zowel afzonderlijk als in combinatie getest worden? • Is in de technische documentatie beschreven welke methode(s) is/zijn gebruikt? • Is de gehanteerde analysemethode uitgelegd?

	• Is de nauwkeurigheid van de analysemethode gemeten en beschreven?
Doeltreffendheid en inbedding: Validatiemethoden	• Is er een evaluatiecyclus in plaats gezet om risico's en maatregelen te evalueren? • Is samen bepaald met de materiedeskundigen hoe de voorspelkracht van het model beoordeeld gaat worden? Met welke metrieken en bijbehorende drempelwaarden? • Wanneer is het model goed genoeg? Zijn *false positives* bijvoorbeeld minder erg dan *false negatives*, of is het beide even belangrijk? Hoe kunnen we dit kwantificeren? • De nauwkeurigheid van de analysemethode is gemeten en beschreven.
Doeltreffendheid en inbedding: Parametrisering	• Is de wijze waarop parametrisering tot stand is gekomen beschreven? • Is vastgelegd welke parameters gebruikt zijn? • Is geanalyseerd en beschreven wat de gevolgen zijn van een andere parameterkeuze voor de resultaten?

Tabel 1: Evaluatieaspecten en verwijzing naar de specifieke vragen t.a.v. auditeerbaarheid en doeltreffendheid en inbedding, vanuit de richtlijnen van Ministerie van Justitie en Veiligheid [7, 18].

3.3 Technische analyse van data, code en uitkomsten

TDL en Nissewaard hebben voor dit onderzoek de data en code van het AI-algoritme gedeeld met TNO. In de technische analyse die TNO hiermee heeft uitgevoerd, zijn een aantal onderwerpen en specifieke vragen afgeleid uit de richtlijnen van MinJenV (zie Sectie 4.1) en deze zijn vervolgens beantwoord. De resultaten hiervan zijn te vinden in Sectie 4.2.

Auditeerbaarheid*:* Auditeerbaarheid van een algoritme betreft de mate waarin het mogelijk is om te controleren of het algoritme aantoonbaar voldoet aan inhoudelijke en procedurele eisen. De werkwijze om auditeerbaarheid te onderzoeken is gebaseerd op data- en code-inspectie:

- *herhaalbaarheid*: hierbij wordt onderzocht of bij het meermaals draaien van de code op dezelfde data, in dit geval door TDL, steeds dezelfde uitkomsten worden geproduceerd.
- *reproduceerbaarheid*: hierbij wordt geverifieerd of de uitkomsten of suggesties van TDL kunnen worden gereproduceerd door een andere partij, in dit geval TNO.
- *versiebeheer*: voor versiebeheer is een inkijk gegeven door TDL en zijn vervolgvragen geformuleerd aan de hand van de richtlijnen en TNO-expertise.
- *technische documentatie*: Voor technische documentatie is zowel de documentatie in de code als de documentatie over de code geïnspecteerd en beoordeeld.

Doeltreffendheid en inbedding: Onder doeltreffendheid wordt verstaan de mate waarin door het algoritme aangeduide (niet-)risicovolle cliënten, overeenkomen met de daadwerkelijk geconstateerde risico's waarvan na menselijke controle is vastgesteld dat oneigenlijk gebruik of misbruik (niet) heeft plaats gevonden.
De focus van het onderzoek ten aanzien van doeltreffendheid van het algoritme in haar inbedding is gelegd op de robuustheid van de uitkomsten van het algoritme. Deze robuustheid is hoog als een algoritme dezelfde kwaliteit behoudt bij veranderingen in de invoerdata en niet afhankelijk is van de toevallig getrokken dataverzameling (*sampling*) waarop het algoritme wordt getraind. Bij het door TDL ontwikkelde algoritme wordt bij het trainen gebruik gemaakt van *sampling* van data. Beiden zijn bedoeld om het algoritme robuust te maken, maar introduceren mogelijk ook afhankelijkheid van de specifieke samenstelling van de voor training getrokken dataverzameling. Hierdoor kan ook de opgeleverde uitkomsten, in dit geval een

rangschikking van cliënten op basis van risicoscores, worden beïnvloed door de willekeurige trekkingen in het leeralgoritme.

Andere onderzochte aspecten, die gerelateerd zijn aan de data en code kwaliteit, zijn de testprocedures, validatiemethoden en parametrisering:

- *Robuustheid:* Het onderzoek naar robuustheid heeft als doel om aan te geven in welke mate de suggesties van het algoritme afhankelijk zijn van de willekeurige trekkingen door het algoritme. Meer specifiek is dit onderzocht door de analyse meerdere keren op dezelfde data uit te voeren en de suggesties van het algoritme te bestuderen die bepaald worden door de rangschikking van de risicoscores. Bij iedere herhaling krijgt elke cliënt van de dataset een ranking, waarbij de cliënt met de hoogste risicoscore een ranking van 1 krijgt en de cliënt met de één na hoogste risicoscore een ranking van 2, enzovoort. Een robuust algoritme geeft een cliënt over meerdere herhalingen een soortgelijke ranking: een cliënt krijgt bijvoorbeeld elke herhaling een ranking in de top-10 (b.v. eerst ranking van 3, dan van 4, daarna van 3, enz.). Een niet-robuuste algoritme geeft een cliënt sterk verschillende rankings (b.v. een cliënt krijgt eerst een ranking van 3, maar bij de volgende herhaling een ranking van 1201, en daarna een ranking van 272, enz.). Om te onderzoeken of het algoritme robuust genoeg is om op basis van één uitvoering suggesties te doen voor vervolgonderzoek is gekeken naar drie aspecten: (1) hoe vaak komen de cliënten in de top-10 van de eerste analyse nogmaals in de top-10 van de 100 herhalingen, (2) de gemiddelde ranking per cliënt over 100 herhalingen en (3) de spreiding van de ranking per cliënt over 100 herhalingen. De nadruk is gelegd op de robuustheid van de top-10, omdat in de huidige inbedding Nissewaard voor de cliënten in deze top-10 een vervolgonderzoek kan/zal opstarten.
- *Testprocedures:* Er heeft inspectie van de door TDL uitgevoerde testprocedures plaatsgevonden om inzicht te verkrijgen in hoe de doeltreffendheid in de code van het algoritme wordt gemeten.
- *Validatiemethoden:* De door TDL en Nissewaard gebruikte validatiemethoden zijn onderzocht om inzicht te verkrijgen in hoe de doeltreffendheid van het algoritme in zijn inbedding wordt gemeten.
- *Parametrisering*: Aan de hand van code-inspectie is de procedure om de parameters van het model te kiezen onderzocht.

4 Resultaten van de technisch-inhoudelijke evaluatie

In dit hoofdstuk presenteren we de resultaten van de technisch-inhoudelijke evaluatie.

4.1 Resultaten van het combineren van richtlijnen

De richtlijnen van MinJenV [7, 18] geven handvatten voor de ontwikkeling en het gebruik van algoritmen door de overheid en ten behoeve van de publieksvoorlichting. Deze richtlijnen zijn als kamerstuk openbaar gesteld. De richtlijnen bevatten de volgende onderwerpen:

- Bewustzijn en inperking van risico's;
- Uitlegbaarheid;
- Gegevensherkenning;
- Auditeerbaarheid;
- Verantwoording;
- Validatie;
- Toetsbaarheid;
- Publieksvoorlichting.

De richtlijnen van EC [9] geven een kader voor het bewerkstelligen van betrouwbare AI (oftewel: *trustworthy AI*). Deze richtlijnen zijn gebaseerd op vier ethische beginselen: (1) respect voor menselijke autonomie, (2) preventie van schade, (3) rechtvaardigheid en (4) verantwoording. Uit deze beginselen zijn zeven vereisten geformuleerd:

- Menselijke controle en menselijk toezicht;
- Technische robuustheid en veiligheid;
- Privacy en data governance;
- Transparantie;
- Diversiteit, non-discriminatie en rechtvaardigheid;
- Maatschappelijk en milieuwelzijn;
- Verantwoording.

In het formuleren van de vragenlijst is, in de vergelijking tussen beide richtlijnen, gekozen om uit te gaan van de richtlijnen van MinJenV, aangevuld met een aantal vragen uit de richtlijnen van EC. De belangrijkste redenen voor de keuze van de richtlijnen van MinJenV zijn als volgt:

(1) De richtlijnen MinJenV sluiten beter aan bij de gekozen focus van de Nissewaard-casus dan de EC-richtlijnen. Een aantal verschillende EC-richtlijnen vallen buiten de scope van de technische evaluatie van dit project. Denk hierbij aan richtlijnen met betrekking tot milieu-welzijn en security die direct onder de AVG vallen.
(2) De richtlijnen van MinJenV richten zich direct op de doelgroep "overheid", waar gemeenten zoals Nissewaard onder vallen. De richtlijnen van de EC zijn daarentegen eerder gericht op organisaties buiten de publieke sector.
(3) De richtlijnen van MinJenV hebben vragen omtrent technische aspecten welke meer verfijnd zijn dan die van de EC-richtlijnen. Dit sluit aan bij ons doel van een technische evaluatie op het gebruik van het algoritme door Nissewaard.
(4) Uit analyse van recentelijke rechtelijke uitspraken, waaronder de reeds genoemde uitspraak in zake SyRI, blijkt dat Nederlandse rechters eerder refereren aan de ethische uitgangspunten zoals verwoord in de richtlijnen van MinJenV dan aan de richtlijnen van de EC.

De volledige vragenlijst is te vinden in Bijlage 1. Tabel 2 geeft een overzicht van de richtlijnen van MinJenV en EC en de link naar de gekozen technische evaluatieaspecten.

Evaluatieaspecten	Richtlijnen Ministerie J&V (onderdelen van)	Richtlijnen Europese Commissie (onderdelen van)
Auditeerbaarheid	**Auditeerbaarheid** *(Controleerbare validatie, reproduceerbaarheid, onderbouwing)* **Gegevensherkenning** *(Parameterisering)* **Uitlegbaarheid** *(Modulaire code, testbaar, technische documentatie)*	**Accountability** *(Auditability, reporting procedure, redress)* ***Technical robustness and safety*** *(Reproducibility)* **Transparency** *(Explainability, communication)*
Doeltreffendheid en inbedding	**Bewustzijn en inperking risico's** *(Test-procedure, robuustheid)* **Validatie** *(Validatiemethoden)*	**Human agency and oversight** *(Proportionality, subsidiarity)* **Technical robustness and safety** *(Reliability)*
-	*Enkel interviewinzichten over:* ***Verantwoording*** ***Toetsbaarheid*** ***Publieksvoorlichting***	*Enkel interviewinzichten over:* ***Diversity, non-discrimination and fairness*** ***Societal and environmental wellbeing*** ***Privacy and data governance***

Tabel 2: De relatie tussen de evaluatieaspecten uit de technische analyse en onderdelen van de richtlijnen waarop de interviews zijn gebaseerd. In de onderste rij van de tabel is te zien dat een aantal richtlijnen geheel buiten de scope van de projectafbakening vallen. Verder benadrukken we dat de onderdelen van bepaalde richtlijnen zoals "uitlegbaarheid" zijn meegenomen, maar dat uitlegbaarheid niet volledig onder het onderzoeksgebied "auditeerbaarheid" valt.

4.2 Resultaten vanuit de interviews en technische analyse

De eerste inzichten per ethische richtlijn zijn verkregen via meerdere interviews met Nissewaard en TDL. De volledige vragenlijst te vinden in Bijlage 1. De samenvatting van de resultaten van de interviews voor de verschillende richtlijnen zijn te vinden in Bijlage 2. Aanvullend is de technische analyse gedaan.

Deze sectie beschrijft de gecombineerde resultaten van de interviews en de technische analyse per gekozen evaluatieaspect.

4.2.1 Auditeerbaarheid

Interviews

Uit de interviews met Nissewaard en TDL blijkt dat er op meerdere vlakken aan auditeerbaarheid is gewerkt. Zo is oorspronkelijk de samenwerking tussen Nissewaard en TDL gestart met een data-onderzoek (genaamd de '*data deep dive*'), waarbij onder meer gekeken werd naar keuzes omtrent de te gebruiken data en analysemethode. Ook is aandacht besteed aan documentatie, periodieke evaluatie en de uitlegbaarheid van het algoritme. Belangrijke interviewinzichten voor auditeerbaarheid zijn:

Auditeerbaarheid – Herhaalbaarheid, Reproduceerbaarheid en Versiebeheer:

- Het algoritme is gebaseerd op het "*rule-fit*" model dat beschreven is in [16]. De specifieke implementatie en gemaakte keuzes zijn niet openbaar beschikbaar gemaakt door TDL, maar deze zijn wel beschikbaar in zodanige vorm dat een derde partij deze kan evalueren.
- Ook is aangegeven dat de uitkomsten van het algoritme reproduceerbaar zouden moeten zijn en versiebeheer geregeld is.

Auditeerbaarheid – Technische documentatie:

- Het gebruik en keuze van databronnen voor training en testen zijn gedocumenteerd. Zo is in de technische documentatie aangegeven welke variabelen na afwegingen zijn meegenomen. Deze specifieke afwegingen zijn overigens niet opgenomen in deze technische documentatie.
- Er is aangegeven en gedocumenteerd dat er in de ontwikkeling van het algoritme specifiek is gekozen voor een meer uitlegbaar en transparant algoritme, vergeleken met een algoritme dat mogelijk accurater maar minder uitlegbaar is.

Technische analyse

Auditeerbaarheid – Herhaalbaarheid, Reproduceerbaarheid:
Uit onze technische analyse blijkt dat het algoritme van TDL niet tot reproduceerbare noch herhaalbare resultaten leidt.

- Context: het algoritme van TDL is niet deterministisch, dat wil zeggen dat er elementen van willekeurigheid (ook wel: *randomness*) in het algoritme zitten. Dit hoeft op zichzelf geen probleem te zijn en komt vaak voor bij bepaalde typen AI-algoritmen. Echter, voor reproduceerbaarheid van een algoritme is het van belang dat de willekeurige elementen opnieuw “op exact dezelfde manier willekeurig” kunnen worden uitgevoerd. Dan kan bijvoorbeeld de gebruiker van een algoritme-toepassing altijd exact herhalen hoe een bepaalde uitkomst is berekend en zich verantwoorden.
- Bevinding: deze mogelijkheid om de willekeurigheid van een analyse te herhalen is niet uitgevoerd en niet opgeslagen door TDL.[3] Om deze reden voldoet het algoritme van TDL niet aan de aspecten herhaalbaarheid (oftewel: reproduceren van uitkomsten met dezelfde data en code door dezelfde gebruiker) en reproduceerbaarheid (oftewel: reproduceren van uitkomsten met dezelfde data en code door een andere gebruiker).
- De code-review heeft zwakheden blootgelegd ten aanzien van de consistentie, de noodzaak tot handmatige handelingen en leesbaarheid: De configuratie is niet volledig beschreven aan het begin van het script, maar staat verspreid over de code.
- Handmatige handelingen, diep in de code, zijn nodig om verschillende functionaliteiten van het script te kunnen uitvoeren.
- De code is niet gemakkelijk leesbaar voor software-ontwikkelaars of data-wetenschappers (oftewel: *data scientists*) die niet bij de totstandkoming waren betrokken. Hierbij speelt o.a. dat geen *best coding practices* zijn toegepast, zoals consistente en duidelijke naamgeving van data-elementen.

Auditeerbaarheid – Versiebeheer:

- Een inkijk in versiebeheer bij TDL heeft laten zien dat versiebeheer is geregeld door gebruik te maken van GitHub. Zo worden aanpassingen aan de code bijgehouden, inclusief datum en *comments*.
- TDL heeft aangegeven dat zij databestanden, testresultaten, parameters en uitkomsten opslaat voor elke ‘data-oplevering’ aan Nissewaard.

Auditeerbaarheid – Technische documentatie:
Uit de code-review is gebleken dat de technische documentatie niet voldoende gedetailleerd is voor de behoeften van de softwarebeheerder/gebruiker en softwareontwikkelaars.

- Dit is een belemmering om overzicht te krijgen van de code en zijn functionaliteiten. Bovendien vermindert dit de auditeerbaarheid van het algoritme. Meer toelichting is te vinden in de aanbevelingen in Hoofdstuk 5.

[3] Voor de technische lezer, een veelgebruikte manier om de willekeurigheid vast te leggen is door gebruik en opslaan van een “*seed*” voor de “*random number generator*”.

4.2.2 Doeltreffendheid en Inbedding

Interviews

Uit de interviews is gebleken dat in de aanpak van Nissewaard veel aandacht is besteed aan de inbedding van het algoritme en overleg met de verschillende belanghebbenden. Bijstandsontvangers zijn daarbij vertegenwoordigd door de Landelijke Cliëntenraad. Er lijkt tot nu toe echter minder aandacht besteed te zijn van het meten van de doeltreffendheid en de validiteit van het algoritme zelf. Belangrijke bevindingen die naar aanleiding van de interviews zijn geconstateerd met betrekking tot over de doeltreffendheid en inbedding zijn hieronder beknopt te vinden, zie Bijlage 2 voor meer toelichting.

Inbedding:

- Ter voorbereiding van de algoritmetoepassing heeft Nissewaard vooronderzoek gedaan bij o.a. andere gemeenten, de onafhankelijke functionaris gegevensbescherming van Nissewaard en Stimulansz (een stichting die als kennis- en adviespartner fungeert van gemeenten). In samenspraak met de verschillende belanghebbenden is daarna voor de huidige inbedding gekozen.
- Er is veelvuldig gecommuniceerd met de gemeenteraad over de toepassing van het algoritme. Het model is in 2019 op congressen gepresenteerd aan andere gemeenten (zoals het jaarlijkse congres Overheid 360° [23]) en was onderwerp van onderzoek op het *Transparency Lab*, georganiseerd door ministerie BZK, zie [22]. Het model is verder door de Universiteit Utrecht aan een scan van De Ethische Data Assistent (DEDA) onderworpen [20].
- De gebruikers van de uitkomsten van het algoritme, de sociaal rechercheurs bij Nissewaard, zijn meegenomen in het proces van de ontwikkeling en evaluatie van het algoritme. Bovendien ontvangen zij jaarlijks een opfriscursus rondom het algoritme.
- Het algoritme is een aanvullend instrument om misbruik en oneigenlijk gebruik van bijstand tegen te gaan. De uitkomsten van het AI-model worden alleen gebruikt als een suggestie, en de top-10 resultaten worden onderzocht door rechercheurs van Nissewaard. Suggesties van het algoritme maken 10-15% van het werk van de sociaal rechercheurs uit; het merendeel bestaat uit het nalopen van meldingen (van b.v. een buurman die verdacht gedrag opmerkt). De gehele ranking wordt overigens wel aan Nissewaard geleverd door TDL. Nissewaard geeft aan dat niet is gekozen voor tussenvormen van top-50 of top-100, wegens beperkte capaciteit van de sociaal rechercheurs.
- Bij elke data-oplevering van Nissewaard aan TDL wordt door TDL-onderzoek gedaan naar veranderingen in de aangeleverde data in termen van koppelbaarheid, variatie, vulling en bias (voor elke variabele). Afwijkingen worden besproken met de gemeente.
- Het AI-algoritme van TDL leert onderscheid te maken tussen cliënten waarvan is vastgesteld dat sprake was van bijstandsmisbruik of oneigenlijk gebruik en alle overige cliënten. Informatie over suggesties van het algoritme die na controle toch geen misbruik of oneigenlijk gebruik betreffen (zogeheten “valspositieven”) wordt niet meegenomen in het (her)trainen van het algoritme.

Doeltreffendheid – Robuustheid:

- TDL geeft aan dat voor robuustheid bij elke ontvangst van nieuwe data, het model en de uitkomsten worden gecontroleerd op beïnvloeding door enige veranderingen in de data. Waar nodig wordt het algoritme verbeterd, aanpassingen worden vastgelegd, en ook onverwachte resultaten worden door TDL vastgelegd en besproken in een evaluatiesessie met Nissewaard.

Doeltreffendheid – Testprocedures:

- De nauwkeurigheid van het algoritme wordt geanalyseerd door TDL door middel van een “training en test split” zoals omschreven in Sectie 2.2. In de code voor de test-procedure wordt er één splitsing gemaakt tussen training (75% van de data) en testing (25% van de data).
- Op basis van een AI-model getraind op het training-gedeelte van de data worden voorspellingen gedaan voor het test-gedeelte van de data. De voorspellingen worden vooral getoetst aan de hoeveelheid suggesties in de top-10 en top-100 ook bewezen gevallen betreffen. Voor de meer technische lezer: het AI-model wordt verder geëvalueerd met o.a. een *ROC curve*, *Importance curve* en de *Mean Squared Error* [1].

Doeltreffendheid – Validatie:

- De doeltreffendheid van het algoritme in het vigerende toezichtproces wordt vergeleken met de oorspronkelijke invulling van het handhavingsproces. Daarbij diende bijna iedereen die bijstand ontvangt bij een jaarlijkse controle verklaringen af te geven en werd iedere uitkeringsontvanger onderzocht. In de huidige opzet van het proces zijn deze jaarlijkse controles vervangen door risicogerichte handhaving. Dit brengt aanzienlijk minder lasten bij de bijstandsontvangers teweeg. De beschikbare controlecapaciteit wordt nu ingezet voor een diepgaandere controle op een aanzienlijk kleiner aantal cliënten.
- Door Nissewaard is aangegeven dat de doeltreffendheid van het algoritme in de inbedding van het handhavingsproces fluctueert en rond de 2 à 3 successen in de top-10 ligt. Tot dusver werd door Nissewaard de validiteit van het algoritme vooral afgezet tegen de vorige invulling van het handhavingsproces, waarbij ruim 2300 huishoudens een vervolgonderzoek kregen dat minder diepgaand was. Hierbij wordt bij het bepalen van de validiteit van het algoritme dus vooral gekeken naar de effecten als gevolg van de verandering om in plaats van iedere bijstandsontvanger een kleine selectie te onderzoeken.
- Daarnaast is in het proces niet afgesproken wat de drempelwaarden van het algoritme moeten zijn en of bijvoorbeeld “vals positieven” (foute suggesties) minder erg zijn dan “valsnegatieven” (gemiste suggesties).
- De analysemethode waarmee de nauwkeurigheid van het algoritme is vastgesteld is nader beschreven in de technische documentatie [1].

Doeltreffendheid – parametrisering:

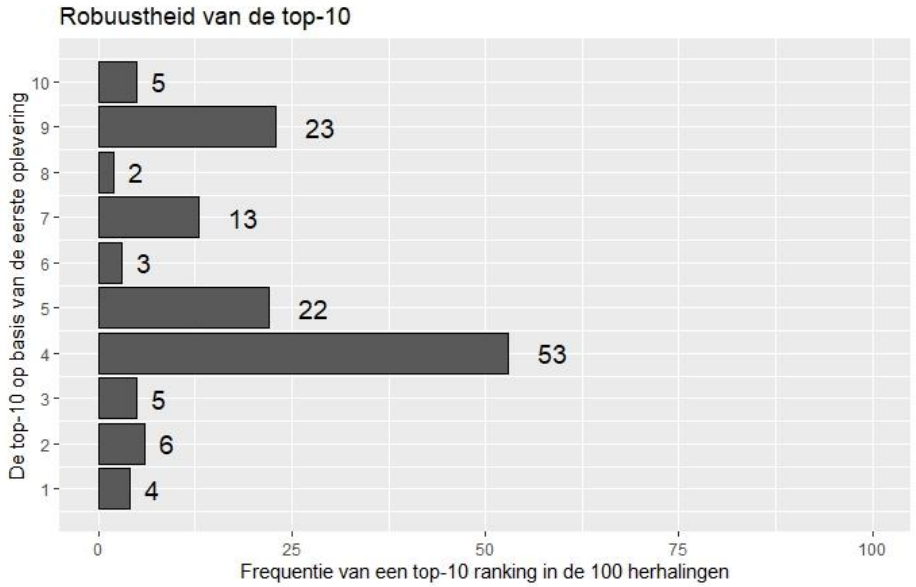
- In de parametrisering is na de keuze voor de trainingsgegevens een *rule-fit* AI-model getraind, waarbij de parameters (diepte van de boom, aantal bomen en de zogeheten lambda-parameter in het AI-model) zijn afgestemd. Dit is beschreven in de technische documentatie [1], zie ook [17].

Technische analyse

Robuustheid:

Uit de technische analyse blijkt dat de mate van robuustheid van het algoritme van TDL laag is. Omdat de resultaten van dit algoritme niet buiten de TDL-setting kunnen worden gereproduceerd, zijn de voor het meten van robuustheid noodzakelijke herhalingen uitgevoerd bij en door TDL[4].

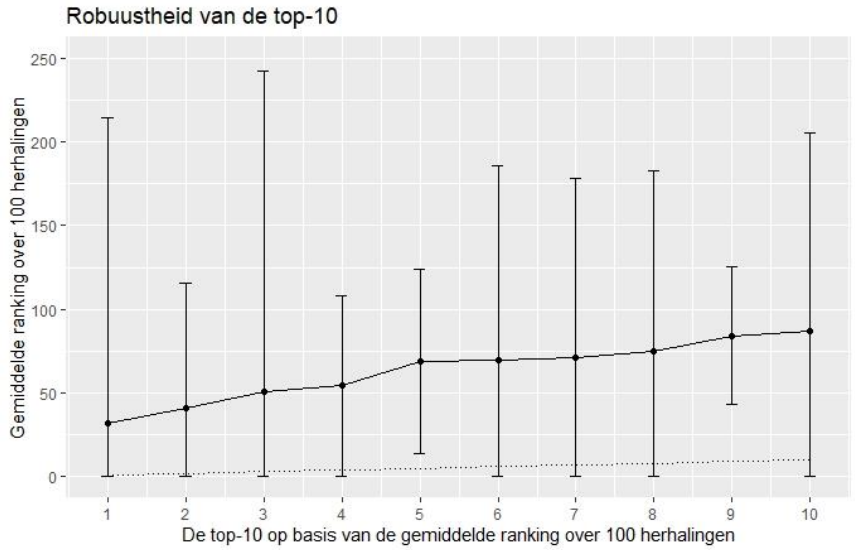
Het gebrek aan robuustheid van het algoritme maakt dat de doeltreffendheid van het risicogestuurd handhaven niet kan worden bepaald op basis van eenmalige toepassing van het algoritme. Hieronder staan de resultaten van experimenten op basis van 100 herhaalde toepassingen van de training- en opleverprocedure.



Figuur 4: De robuustheid van de uitkomsten in de huidige inbedding is aangegeven aan de hand van de cliënten die in de eerste herhaling zouden worden gesuggereerd voor verder onderzoek. Van deze cliënten is aangeduid hoe vaak ze zouden worden gesuggereerd aan de hand van dezelfde data en code in de 100 herhalingen. Dit geeft aan hoe vaak de top-10 cliënten van de eerste herhaling ook voorkomen in de top-10 van de andere 99 herhalingen.

[4] De specificaties om deze herhalingen te reproduceren zijn als volgt opgegeven door TDL.

- Computervereisten: Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz (10 processors), 48 GB RAM en 64-bit Operating System.
- Data-versie: “40129 Nissewaard Oplevering 2021”, dat is de data van 24 februari 2021.
- Code-versie: 8 januari 2021.
- Software-versie: R version 3.6.2, versies van R-packages zijn ook bekend.
- Oplevering peildatum: 22-12-2020.
- Seed: 12345.



Figuur 5: De robuustheid van de uitkomsten is aangegeven aan de hand van de gemiddelde ranking en de verdeling van de rankings per client met de doorgetrokken zwarte lijn. De banden geven hier een standaarddeviatie aan boven en onder het gemiddelde van honderd herhalingen. Waar de ondergrens een waarde onder nul bevat, is hier gekozen om deze als nul te illustreren, gegeven dat een negatieve ranking hier niet mogelijk is. Met een stippellijn is een benchmark afgebeeld van een ideale niet-willekeurige ranking, die elke herhaling consistent dezelfde ranking geeft per client.

Interpretatie van de figuren:

- Figuur 4 geeft een intuïtief beeld over de robuustheid van de top-10 suggesties in de context van de huidige inbedding van het algoritme. In deze huidige inbedding geeft TDL aan Nissewaard een ranking van cliënten gebaseerd op de eerste uitvoering van het algoritme op de data. Dit is ook weergeven in het staafdiagram in Figuur 4, waar de verticale y-as de top-10 cliënten aanduidt die op basis van de eerste ranking voor vervolgonderzoek worden gesuggereerd. De horizontale staven geven aan hoe vaak de cliënten in de eerste top-10 van alle 100 herhalingen bij de top-10 suggesties toebehoren. We zien dat de meeste cliënten in de eerste top-10 minder dan 10 keer in de top-10 voorkomen in de 100 herhalingen en dat niet één cliënt in alle 100 herhalingen in de top-10 zit. Voor risicogestuurd handhaven op basis van één uitvoering is robuustheid belangrijk. Hoge maximale robuustheid zou zich in Figuur 4 uiten met frequenties van de eerste top-10 dichtbij 100.
- Figuur 5 geeft inzicht in de robuustheid van het algoritme zonder direct de huidige inbedding in acht te nemen, waar suggesties op basis van één uitvoering worden gegeven. De y-as geeft de gemiddelde ranking over 100 herhalingen weer en de x-as geeft aan dat we dit gemiddelde laten zien van de top-10 cliënten met de laagste gemiddelde ranking over 100 herhalingen, oftewel: de top-10 cliënten die door het algoritme het vaakst voor vervolgonderzoek zouden zijn gesuggereerd vanwege hun risicoscore. De doorgetrokken lijn laat zien dat de cliënten met de twee laagst gemiddelde rankings een gemiddelde hebben van 32 en 41. De verticale zwarte lijnen in Figuur 5 geven per cliënt de spreiding van de ranking door middel van één standaarddeviatie boven en onder het gemiddelde. De standaarddeviaties per cliënt zijn hoog, waarvan de kleinste standaarddeviatie 41 rankings betreft. De stippellijn van (0,0) tot (10,10) is het resultaat van een algoritme dat altijd dezelfde resultaten oplevert. Dit is als benchmark toegevoegd om de robuustheid van het algoritme van TDL te vergelijken hiermee. De gemiddelde ranking van de top-10 die duidelijk boven een ranking van 10 liggen en de hoge standaarddeviaties geven aan

dat de rankings van de top-10 cliënten substantieel variëren tussen herhalingen.
- Bevinding: de 100 herhalingen van het algoritme hebben laten zien dat de top-10 van één oplevering niet robuust is. Zo laat Figuur 4 zien dat het veel voorkomt dat een suggestie in de top-10 in de ene herhaling niet voorkomt in de top-10 bij een andere herhaling. In de huidige inbedding wordt een cliënt die op basis van één analyse als risicovol geschat, vervolgens verder onderzocht door de sociaal rechercheur. Op dezelfde data met dezelfde code zou dezelfde cliënt in een herhaling niet als risicovol worden geschat. Ook buiten de huidige inbedding van één oplevering geeft het AI-algoritme geen stabiele suggestie voor vervolgonderzoek. Deze onzekerheid is niet verder onderzocht. TDL vangt deze willekeurigheid ook niet af bij het testen van de nauwkeurigheid van het algoritme. Samengevat zijn de door het AI-algoritme aangeduide risico's voor een belangrijk deel willekeurig bepaald, wat ongewenst is ten aanzien van robuustheid.

Testprocedures en validatiemethoden:
- De code laat zien dat er geen tests worden gedaan om de correcte functionaliteit te waarborgen, zowel bij het afzonderlijk runnen van de delen in de code als in combinatie. Het algoritme is wel getest aan de hand van de testdata, maar zoals aangegeven is dit niet herhaaldelijk op dezelfde data gebeurd voorafgaand aan een oplevering.
- In de testprocedures van het algoritme wordt geen rekening gehouden met de verhouding tussen cliëntendossiers met en zonder bewezen misbruik of oneigenlijk gebruik tussen training- en testgedeeltes van de data. Met het feit dat enkel één training- en testsplit wordt gedaan kan een onbalans in de verdeling een substantiële invloed hebben op de gemeten nauwkeurigheid.
- In de testprocedure van het algoritme wordt in de indeling van de training- en testset geen rekening gehouden met de tijdsvolgorde in de toepassing van het algoritme. In de risicogestuurde handhaving van Nissewaard wordt informatie van bijstandsgerechtigden die over een langere periode in het verleden is verzameld gebruikt om het risico van huidige bijstands-ontvangers te voorspellen. Bij het samenstellen van de train- en testset wordt de tijdsdimensie buiten beschouwing gelaten. Cliënten worden willekeurig toegedeeld aan de training- en testset. Het kan dus gemakkelijk voorkomen dat verdeling van data over de training- en testset scheef is ten aanzien van de periode die door de data wordt beschreven.

Parametrisering:
- Inspectie in de code heeft laten zien dat in de testprocedure van de code verschillende parameters bekeken kunnen worden, inclusief interacties tussen verschillende parameters. Verschillende waarden voor de parameters worden door TDL handmatig ingevuld midden in het script in plaats van in een configuratiedocument of boven in het script en de resultaten worden manueel vergeleken door middel van prints van de resultaten. Hierdoor is het lastig om met verschillende tests te controleren welke parameters zijn gebruikt. Het afstemmen van de parameters gebeurt per datalevering, waarbij handmatig een aantal parameterkeuzes vergeleken worden.
- Een uitgebreide analyse waarbij de invloed op de uitkomsten van een andere parameterkeuze wordt onderzocht is niet inbegrepen in de test-procedure van de code. Zo is niet bekend of een andere parameterkeuze andere suggesties oplevert en op grond waarvan de betreffende parameterinstelling gekozen is.

5 Conclusies en aanbevelingen

TNO heeft een technisch-inhoudelijke evaluatie uitgevoerd van het AI-algoritme dat door Gemeente Nissewaard wordt gebruikt voor het opsporen van misbruik en oneigenlijk gebruik van bijstandsuitkeringen. Dit algoritme is ontwikkeld en geïmplementeerd door Totta Data Lab. In de voorgaande hoofdstukken zijn de globale aanpak, de evaluatiewerkwijze en de evaluatieresultaten gepresenteerd. Hieronder staan de hoofdbevindingen, conclusie en aanbevelingen.

5.1 Hoofdbevindingen

De positieve bevindingen ten aanzien van (de inzet van) het AI-algoritme zijn als volgt:

- In het huidige protocol van de Gemeente Nissewaard worden de risico-indicaties die door het AI-algoritme worden aangegeven gecombineerd met andere risico-indicaties, waarbij de door het AI-algoritme aangedragen te onderzoeken dossiers slecht 10-15% van de te onderzoeken populatie bedraagt. Deze risico-indicaties spelen een rol in de selectie van dossiers, maar hebben verder geen sturende invloed op het verder handmatig uitgevoerde toezicht- en handhavingsproces. De mens draagt en neemt vervolgens ook de verantwoordelijkheid voor mogelijke besluiten die uit het toezicht- en handhavingsproces voortvloeien.
- Uit interviews is gebleken dat zowel de Gemeente Nissewaard als Totta Data Lab aandacht heeft besteed aan verschillende aspecten van verantwoord gebruik van AI-algoritmen, zoals datagebruik, non-discriminatie, transparantie, verantwoording, uitlegbaarheid en publieksvoorlichting (beschreven in de richtlijnen). Naast de interviews zijn deze aspecten niet verder onderzocht in de technische analyse van de evaluatie.

Op de volgende punten presteert het AI-algoritme onvoldoende:

- De auditeerbaarheid van het AI-algoritme is gebrekkig. Dit betekent dat het algoritme in de huidige vorm niet voldoende toetsbaar is aan inhoudelijke en procedurele eisen. De uitkomsten van het algoritme blijken namelijk noch herhaalbaar noch reproduceerbaar te zijn, omdat er onvoldoende rekening is gehouden met de willekeurigheid die inherent is aan het algoritme. Dit betekent dat bij meerdere runs van dezelfde code met dezelfde instellingen er (significant) andere resultaten uitkomen. Daarnaast is de kwaliteit van de programmeercode niet op peil. Bij het implementeren van het algoritme, maar ook bij het testen van de code, zijn de *best practices* vanuit software-ontwikkeling onvoldoende toegepast. Evenmin ontbreekt het aan aanvullende documentatie voor softwareontwikkelaars en technische gebruikers van de code. Aspecten zoals reproduceerbaarheid en kwaliteit van de code zijn een voorwaarde voor het kunnen auditeren van andere relevante ethische aspecten, zoals non-discriminatie en transparantie van een algoritme-toepassing. Tijdens de interviews is geconstateerd dat er wel aandacht is besteed aan deze andere aspecten, maar deze konden niet verder worden onderzocht vanwege het gebrek aan auditeerbaarheid.
- Het is onvoldoende mogelijk om de doeltreffendheid (ook wel: effectiviteit) van het AI-algoritme vanuit een technisch oogpunt te onderzoeken. De oorzaak is dat de resultaten niet robuust zijn met betrekking tot de aanwezigheid van de willekeurige elementen in het algoritme. Dit was ook de oorzaak van het bovengenoemde probleem ten aanzien van herhaalbaarheid en reproduceerbaarheid van het algoritme. Uit de

evaluatie blijkt dat de met behulp van het AI-algoritme vastgestelde risicoscores, inclusief de berekende volgorde van cliënten op basis van dergelijke risicoscores, niet betrouwbaar zijn. Hierdoor kan Nissewaard de uitkomsten van het algoritme niet met een gepast niveau van zekerheid interpreteren.

5.2 Hoofdconclusie

Ondanks dat Nissewaard en Totta Data Lab duidelijk aandacht hebben besteed aan verschillende aspecten van verantwoord gebruik, is de hoofdconclusie van de technisch-inhoudelijke evaluatie dat het AI-algoritme in de huidige vorm van onvoldoende niveau is om verantwoord te kunnen inzetten.

Deze hoofdconclusie wijst uitsluitend op de huidige vorm van dit specifieke AI-algoritme. TNO velt nadrukkelijk geen oordeel over juridische aspecten of op principiële bezwaren tegen de inbedding van een AI-algoritme in het huidige informatiegestuurde handhavingsproces dat door de Gemeente Nissewaard wordt gehanteerd.

In de hiernavolgende aanbevelingen stelt TNO een aantal stappen voor, die kunnen leiden tot verbetering van de inzet van het AI-algoritme.

5.3 Aanbevelingen

De aanbevelingen vanuit de technisch-inhoudelijke evaluatie richten zich voornamelijk op het verbeteren van (de inzet van) het AI-algoritme en de bijbehorende code.

<u>Aanbevelingen ten aanzien van de methodiek van het AI-algoritme</u>

- De situatie van cliënten kan in de loop der tijd veranderen, waardoor mogelijk nieuwe risico's ontstaan voor misbruik of oneigenlijk gebruik van de bijstand. Op dit moment worden dergelijke statusveranderingen echter niet meegenomen. We raden aan om bij het opnieuw ontwikkelen van een risico-indicatie-algoritme deze statusveranderingen ook mee te nemen. Het is dan ook mogelijk dat het algoritme leert welke soort statusveranderingen en niet leert welke soort cliënten een risico vormen.
- In de opzet van de train- en testset van het AI-algoritme raden we aan om:
  - in de training-test-split ook de tijdsvolgordelijkheid mee te nemen. In toepassing van het algoritme wordt data tot op het heden gebruikt om bewezen gevallen van misbruik en oneigenlijk gebruik in de toekomst (na vervolgonderzoek) te voorspellen. Deze tijdsvolgordelijkheid dient ook in acht te worden genomen in de testprocedure in de code waar de nauwkeurigheid wordt gemeten.
  - rekening te houden met de verhouding van bewezen en overige gevallen tussen de train- en testset. Dit kan bijvoorbeeld proportioneel verdeeld worden wat in meer technisch jargon wordt aangeduid als '*stratified sampling*' van training- en testset.
- De onzekerheden ten aanzien van de predictieresultaten dienen gekwantificeerd te worden, bijvoorbeeld door de gemiddelde ranking en de variatie daarin te rapporteren in aanvulling op de resultaten zoals in dit rapport voorgesteld.
- Criteria voor het vaststellen van *tuning-parameters* dienen te worden vastgelegd en de keuze dient te worden opgeslagen.

<u>Aanbevelingen ten aanzien van de code van het AI-algoritme</u>

- Verbeteren van de testmethodiek.

  - Het toevoegen van een automatische testbank is noodzakelijk om hiermee de werking van o.a. de functies en stukken codes te verifiëren, maar ook om de correctheid van de programmatuur te waarborgen bij updates van de code. Op dit moment lijken er geen of onvoldoende testen te zijn voor de losse functies/modules alsmede om de gehele code te valideren.
  - In ieder geval is het van belang ervoor te zorgen dat het algoritme uitgebreid is getest, voordat de code in de praktijk gebruikt wordt.
- Verbeteren van de software-configuratie.
  - Meer consistentie is noodzakelijk t.a.v. de configuratie van de code. Dit houdt in dat bijvoorbeeld alle paden die op meerdere plekken in de code gebruikt worden in een aparte configuratiefile dienen te worden bijgehouden.
  - Verder is het wenselijk om voor de parameters en data-attributen een heldere naamgeving te hanteren en de configuratieparameters in de handleiding voor de technisch gebruiker op te nemen.
- Verbeteren van de documentatie.
  - Aanvullende documentatie voor software-beheerder/gebruiker en een documentatie voor software-ontwikkelaar is vereist, naast de bestaande (publieke) technische documentatie die meer bedoeld is voor de eindgebruikers en burgers).
  - T.a.v. de documentatie voor software-beheerder/gebruiker gaat het om de exacte beschrijving van het gebruikte model, data, aannames, functies, afhankelijkheden, variabelen en motivatie van de gemaakte keuzes.
  - T.a.v. de documentatie voor *softwareontwikkelaar* gaat het om de lijst van gebruikte tools en een beschrijving van de technische omgeving, hoe de code gecompileerd en gebruikt dient te worden (inclusief voorbeelden), hoe de code is en kan worden getest, en een proces/architectuuroverzicht van de code.

<u>Overige aanbevelingen</u>

- Doen van pilots waarbij het gebruik van het AI-algoritme vergeleken wordt met de oorspronkelijke manier en andere alternatieve manieren van werken, voordat het AI-algoritme daadwerkelijk in productie wordt genomen. Dit geeft inzicht of het gebruik van het algoritme het minst ingrijpende of minst risicovolle alternatief is om het doel te bereiken (het gaat hierbij om subsidiariteit).
- Zodra de bovenstaande aanbevelingen t.a.v. het AI-algoritme en de bijbehorende code zijn opgevolgd, kunnen de andere relevante technische onderzoeksgebieden t.a.v. de betrouwbare AI verder bekeken worden, zoals de exacte keuzes in het gebruik van de data, bias van het algoritme en resultaten en transparantie van het algoritme.
- Blijven evalueren van het AI-algoritme in de toekomst, omdat de ontwikkeling rondom de AI-aspecten datagebruik, doeltreffendheid, transparantie en bias en de technische/ethische/juridische kaders eromheen zich blijven ontwikkelen in de tijd.
- Vaststellen wat de eisen en randvoorwaarden zijn waaraan moet worden voldaan, alvorens een nieuw AI-algoritme te ontwikkelen en te gebruiken: voor welke populatie en onder welke omstandigheden is het model van toepassing, hoe gaat de voorspelkracht beoordeeld worden (inclusief metrieken en drempelwaarden), wanneer is het model goed genoeg, wanneer is er sprake van ongewenste discriminatie (ook hier inclusief metrieken en drempelwaarden), en wanneer van gewenste bias.
- De regelgeving en richtlijnen rond AI-toepassing zijn op nationaal en Europees niveau sterk in ontwikkeling. Het is raadzaam om deze ontwikkelingen te blijven volgen, om aan te sluiten bij de laatste inzichten en richtlijnen rond betrouwbare AI.

6 Referenties

[1] Technische documentatie voorspelmodel bijstandsfraude, Totta data lab, 2019, link: https://nissewaard.notubiz.nl/document/8567516/1/Beantwoording_LOB%2C_SyRi-uitspraak_voor_algoritme_Totta_Data_Lab%2C_bijlage.

[2] Factsheet: Missies voor de Toekomst, inclusief missie overzicht, Nederlandse Topsectoren, 14-11-2019, link: https://www.topsectoren.nl/missiesvoordetoekomst/documenten/publicaties/2019-publicaties/september-2019/23-09-19/factsheet-missies-voor-de-toekomst.

[3] FNV op ramkoers met Nissewaard over fraude-opsporing: rechtszaak dreigt, AD, 2020, link: https://www.ad.nl/voorne-putten/fnv-op-ramkoers-met-nissewaard-over-fraude-opsporing-rechtszaak-dreigt~a4bd9b0c/.

[4] 'Rookgordijn' rondom fraudesysteem Nissewaard, Binnenlands bestuur, 2020, link: https://www.binnenlandsbestuur.nl/sociaal/nieuws/rookgordijn-rondom-fraudesysteem-nissewaard.13026523.lynkx.

[5] Nissewaard belooft burger transparantie over 'fraudescore', FNV, 2020, link: https://www.fnv.nl/nieuwsbericht/sectornieuws/uitkeringsgerechtigden/2020/06/fnv-sceptisch-nissewaard-belooft-burger-transparan.

[6] Raad van State, "Ongevraagd advies over de effecten van de digitalisering voor de rechtsstatelijke verhoudingen," Kamerstukken II 2017/18, 26643, nr. 557, 2018, link: https://www.raadvanstate.nl/@112661/w04-18-0230/.

[7] Bijlage bij "Kamerbrief over waarborgen tegen risico's van data-analyses door de overheid", Rijksoverheid, 8-10-2019, link: https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/08/tk-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid.

[8] SyRI-wetgeving in strijd met het Europees Verdrag voor de Rechten voor de Mens, de Rechtspraak, 5-2-2020, link: https://www.rechtspraak.nl/Organisatie-en-contact/Organisatie/Rechtbanken/Rechtbank-Den-Haag/Nieuws/Paginas/SyRI-wetgeving-in-strijd-met-het-Europees-Verdrag-voor-de-Rechten-voor-de-Mens.aspx.

[9] Ethics guidelines for trustworthy AI, Report, European Commission, 8-4-2019, link: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[10] Fraude opsporen of gevaar van discriminatie? Gemeenten gebruiken 'slimme' algoritmes, NOS.nl, 2021, link: https://nos.nl/artikel/2366864-fraude-opsporen-of-gevaar-van-discriminatie-gemeenten-gebruiken-slimme-algoritmes.html.

[11] Algoritmes zoeken naar bijstandfraudeurs, welke rol speelt etnisch profileren?, NOS.nl, 2021, link: https://nos.nl/artikel/2366962-algoritmes-zoeken-naar-bijstand-fraudeurs-welke-rol-speelt-etnisch-profileren.html.

[12] Ongekend onrecht: verslag – parlementaire ondervragingscommissie kinderopvang toeslag, kamerbrief, 2020, link: https://www.tweedekamer.nl/sites/default/files/atoms/files/20201217_eindverslag_parlementaire_ondervragingscommissie_kinderopvangtoeslag.pdf.

[13] Infographic: Hoe controleren we de uitvoering van de bijstand, gemeente Nissewaard, link: https://www.nissewaard.nl/werk-en-inkomen/werk-en-inkomen_to/infographic-hoe-controleren-we-de-uitvoering-van-de-bijstand.htm.

[14] Uitspraak van de Rechtbank Den Haag in het kort geding t.a.v. e-Screener op 11-02-2020, link: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1013.
[15] Toezicht en handhaving: vangrail van de samenleving! | Whitepaper, VNG Naleving, 20-12-2018, link: https://vng.nl/kennisbank-naleving/toezicht-en-handhaving-vangrail-van-de-samenleving-whitepaper.
[16] Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, *2*(3), 916-954.
[17] Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), European Commission, April 2021, link: https://ec.europa.eu/newsroom/dae/items/709090.
[18] Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses, Ministerie van Justitie en Veiligheid, 1 maart 2021, link: https://led.pleio.nl/files/view/9a19ccc1-65d7-44e7-a86c-c38a2d073eb9/richtlijnen-algoritmen.pdf.
[19] Hoe controleert de gemeente of jij fraudeert?, NRC.nl, 2018, link:https://www.nrc.nl/nieuws/2018/04/06/hoe-controleert-de-gemeente-of-jij-fraudeert-a1598455.
[20] De Ethische Data Assistent (DEDA) van Universiteit Utrecht, link: https://dataschool.nl/deda/.
[21] Uitspraak van de Rechtbank Den Haag in de bodemzaak t.a.v. de SyRI-wetgeving op 05-02-2020, link: https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865.
[22] De bestuurlijke regie over experimentele data en algoritmen, Digitale Overheid, 22 april 2020, link: https://www.digitaleoverheid.nl/led-nieuws/de-bestuurlijke-regie-over-experimentele-data-en-algoritmen/.
[23] Overheid 360°, het jaarlijkse congres over informatiemanagement voor de overheid, link: https://www.overheid360.nl/.

Bijlage 1: Vragenlijst voor interviews t.a.v. de richtlijnen voor betrouwbare AI

In deze bijlage wordt de vragenlijst gepresenteerd die TNO heeft gebruikt voor de Nissewaard-casus. De richtlijnen en teksten van de EC [5] en MinJenV [6, 18] zijn letterlijk overgenomen en gecombineerd en extra vragen zijn toegevoegd. Zoals bediscussieerd in Hoofdstuk 3, is het uitgangspunt van de vragenlijst hierbij de originele richtlijnen van MinJenV [6, 18].

Hieronder wordt in elke sectie een thema van de vragenlijst behandeld, met corresponderende vragen inclusief toelichting en inzichten.

Bewustzijn en inperking van risico's

Bewustzijn en het treffen van maatregelen om de risico's op fouten en ongewenste bias van de toepassing van algoritmen te beperken zodanig dat dit overeenstemming is met de wet.

- Test-cases/scenario's
  - Wordt het algoritme getest op basis van test-cases/scenario's?
  - Vindt dit periodiek plaats?
  - Ook wanneer de software verandert?

- Verandering grootte van en tijdsinvloed op data en de invloed op de uitkomsten.
  - Wordt het algoritme getest wanneer er sprake is van één of meerdere van de onderstaande punten?
    1. Verandering van data
    2. Afname van de data
    3. Toename van de data
    4. Verandering van de dataverdeling
  - Is er een procedure als het algoritme niet voldoet aan het doel van dit algoritme of als er biases in de uitkomsten zijn ontstaan?

- Zorg, binnen de grenzen van de wetgeving[1] voor het opbouwen van controlemechanismes die specifiek toetsen of er geen sprake is van discriminatie of stigmatisering.
  - Wordt er binnen de grenzen van de wetgeving getest of er sprake is van discriminatie of stigmatisering door het algoritme?
    - Uitvoeringswet AVG laat controle met behulp van bijzondere persoonsgegevens, zoals gegevens over iemands ras, vooralsnog niet toe.

- Aanvulling vanuit EC [5]: Oneerlijke bias
  - Hebt u gezorgd voor een strategie of een reeks procedures om te voorkomen dat er onrechtvaardige vertekening wordt gecreëerd of versterkt in het AI-systeem, met betrekking tot zowel het gebruik van inputgegevens als het ontwerp van het algoritme?
  - Hebt u de mogelijke beperkingen die voortkomen uit de samenstelling van de gebruikte gegevenssets, onderzocht en erkend?
  - Hebt u processen ingesteld voor het testen en monitoren van potentiële vertekening gedurende de ontwikkelings-, installatie- en gebruiksfase van het systeem?

  - Hebt u onderzocht of er variabiliteit kan bestaan in de beslissingen die onder gelijke omstandigheden mogelijk zijn?
    - Zo ja, hebt u nagedacht over de mogelijke oorzaken daarvan?
    - In geval van variabiliteit: hebt u een meet- of controlemechanisme vastgesteld voor de potentiële gevolgen van die variabiliteit voor de grondrechten?

- Aanvulling EC [5]: Algoritme-evaluatie ook direct of indirect aan de hand van een concept van rechtvaardigheid.
  - Is er naast het doel van een algoritme om zo effectief mogelijk of zo accuraat mogelijk te zijn ook bewust het concept van “rechtvaardigheid” nagestreefd?
  - Heeft u gezorgd voor een geschikte werkdefinitie van "rechtvaardigheid" die u toepast bij het ontwerpen van AI-systemen?
  - Wordt uw definitie veel gebruikt? Hebt u andere definities overwogen voordat u deze koos?
  - Hebt u gezorgd voor een kwantitatieve analyse of maatstaf om de toegepaste definitie van rechtvaardigheid te meten en te testen?
  - Hebt u mechanismen ingesteld om de rechtvaardigheid in uw AI-systemen te waarborgen? Hebt u andere mogelijke mechanismen overwogen?

- Aanvulling EC [5]: Participatie van belanghebbenden:
  - Hebt u nagedacht over een mechanisme om de participatie van verschillende belanghebbenden onderdeel te maken van de ontwikkeling en het gebruik van het AI-systeem?
  - Hebt u de invoering van het AI-systeem in uw organisatie voorbereid door de getroffen werknemers en hun vertegenwoordigers vooraf in kennis te stellen en bij het proces te betrekken?

- Datakwaliteit en beperkingen. Doel en data noodzakelijk voor doel. Bias en fouten in data.
  - Beschrijf het doel van de analyse.
  - Onderzoek of de data(bron) van voldoende kwaliteit & kwantiteit is[2]:
    1. Is de data relevant voor het doel?
    2. Is de data-aggregatie op een niveau voldoende informatief voor het doel.
    3. Wordt de data over tijd consistent bijgehouden?
    4. Wordt de data op dezelfde manier geregistreerd?
    5. Is er voldoende variatie binnen de data?
    6. Is de data voldoende gevuld?
    7. Zijn de databronnen te koppelen (indien er wordt gewerkt met meerdere tabellen)?
    8. Zijn eventuele mutaties te herleiden? (timestamp e.g.)
    9. Is er een databasewijziging geweest?
    10. Is de manier van dataregistratie gewijzigd?
  - Onderzoek de beperkingen:
    1. Zijn er rechten gevestigd op de databron?
    2. Zijn er rechten gevestigd op het gebruik van het algoritme?
    3. Zijn er rechten gevestigd op het gebruik van de analysemethode?
    4. Welke biases zijn er te vinden in de data?
    5. Welke fouten zijn er te vinden in de data?
    6. Hoe wordt er omgegaan met de biases in de data?
    7. Hoe wordt er omgegaan met eventuele fouten in de data?

- Bewuste keuze voor data-analyse-technieken.
  - Welke technieken zijn getest?
  - Is er een traditionele techniek getest?
  - Waarom is er gekozen om deze specifieke technieken te testen?

  - Waarom is er voor een bepaalde techniek of een combinatie van technieken gekozen?
  - Indien er uiteindelijk voor een kunstmatige-intelligentie-techniek is gekozen, licht toe waarom.
  - Is er gewerkt met vooraf bedachte hypothesen? (voorbedachte regels)

- Standaardmethodiek van werken.
  - Welke standaardmethodiek wordt gebruikt? Beschrijf alle stappen van de modelontwikkeling. Bijvoorbeeld:
    - De data
    - Brondata en typen data
    - Biases & correctie
    - Data-kwaliteit & kwantiteit
    - Omgang met gevoelige gegevens
  - Data verwerken
    - Belangrijke definities
    - Datapreparatie en opschoning
    - Brontabellen werkbaar maken
    - Samenstelling en opbouw van modeltests
    - Basisdataframe
    - Variabelencreatie
    - Datafouten corrigeren
  - De modelontwikkeling
    - Aanpak modelkeuze
    - Uitleg over toegepaste modellen en technieken
  - Resultaten
    - De modelbeoordeling
    - Hoe werkt de modelbeoordeling
  - Gebruik van het algoritme in de praktijk

- DPIA: hanteer gelet op de aard van gegevensverwerking juridische en beleidsmatige toetsingskaders en zorg bij verwerking van persoonsgegevens voor de uitvoering van een DPIA om risico's voor de gegevensbeschermings-rechten van betrokkenen zoveel mogelijk te beperken.
  - Is er een DPIA uitgevoerd?

- Causaliteit.
  - Is het nodig om een causaal algoritme te ontwikkelen om het doel te bereiken?
  - Zo ja/zo nee, waarom?

- Hanteer na implementatie en inrichting van het algoritme een evaluatiecyclus of feedbackloop, zodat wanneer ontwerpers en architecten op afstand komen te staan risico's op tijd worden vastgesteld en adequaat geadresseerd.
  - Is er een evaluatiecyclus in plaats gezet om risico's en maatregelen te evalueren?

- Aanvulling vanuit EC [5]: Inperking van risico autonomie en overmatig vertrouwen.
  - Is er een risico dat het AI-systeem de menselijke autonomie aantast door het beslissingsproces van de eindgebruiker op een onbedoelde manier te beïnvloeden?

- Hebt u voorzorgsmaatregelen genomen om overmatig vertrouwen in of op het AI-systeem bij werkprocessen te voorkomen?
  - Aanvulling EC:

  - Is er afgewogen of door het uitvoeren van het algoritme de data of aannames waarop het algoritme is gebaseerd, ongeschikt kunnen worden? Kan dit tot negatieve gevolgen hebben?
  - Zijn er procedures ingesteld om ervoor te zorgen dat het niveau van nauwkeurigheid van het algoritme dat kan worden verwacht bij de betrokkenen ambtenaren duidelijk is gecommuniceerd?

- Aanvulling vanuit [18]:
  - Is inzichtelijk gemaakt wat het gewicht is van de voorspellende variabelen op de uitkomsten van het model? (oftewel de "*feature importance*")
  - Is inzichtelijk gemaakt hoe de relatie tussen iedere voorspellende variabele en modeluitkomst eruitziet? Is de relatie bijvoorbeeld lineair of non-lineair, en wel of niet monotoon?
  - Is onderzocht of bepaalde voorspellende variabelen zoals buurt of wijk proxy's zijn voor (bijzondere) persoonsgegevens en leiden tot ongewenste discriminatie?
  - Zijn bepaalde deelpopulaties (minderheden) oververtegenwoordigd in de data, wat zou kunnen leiden tot bias in het voorspellend model?
  - Is de data voldoende gevuld en de gebruikte dataset juist, tijdig en volledig?

- Aanvulling vanuit [18]:
  - Zijn er expliciete (kwantitatieve) eisen gemaakt ten aanzien van toepasbaarheid, voorspelkracht, uitlegbaarheid, ongewenste discriminatie, etc., waardoor de oplossingsruimte verder verkleind is?
  - Is aangegeven voor welke populatie en onder welke omstandigheden het model van toepassing moet zijn?
  - Is samen bepaald met de materiedeskundigen hoe de voorspelkracht van het model beoordeeld gaat worden? Met welke metrieken en bijbehorende drempelwaarden? Wanneer is het model goed genoeg? Zijn *false positives* bijvoorbeeld minder erg dan *false negatives*, of is het beide even belangrijk? Hoe kunnen we dit kwantificeren?
  - Wanneer is er sprake van ongewenste discriminatie? Welke metrieken en drempelwaarden worden gebruikt om dit meetbaar en objectiveerbaar te maken?
  - Wanneer is er sprake van ongewenste bias? Is er ook geprobeerd dit te kwantificeren?

Uitlegbaarheid

Uitleg over de uitkomsten en hoe deze tot stand zijn gekomen zijn cruciaal. Uitleg is nodig over het doel, procedures, toegepaste model en algoritme, doorslaggevende variabelen voor de uitkomst en de data. Documenteer al deze aspecten. De inzet van het algoritme moet proportioneel zijn tot het doel. Collegiale uitlegbaarheid, uitlegbaarheid naar derden en technische transparantie. Met als specifieke maatregelen:

- Organiseer de code in modules welke separaat en gecombineerd kunnen worden geëvalueerd.
  - Is er een '*Main*' script waarin alle codes worden aangeroepen? Per gegevensbron is er bijvoorbeeld een code waarin de data is geprepareerd.
  - Is de code in modules georganiseerd?
  - Kan elke module separaat en gecombineerd worden geëvalueerd?

- Test deze modules op correcte functionaliteit zowel afzonderlijk als in combinatie.

  - Zijn alle scripts, waar mogelijk, afzonderlijk van elkaar getest? Zijn alle scripts in combinatie met elkaar getest? Is dit een standaardonderdeel van de modelontwikkeling?
  - Kunnen de modules op correcte functionaliteit zowel afzonderlijk als in combinatie getest worden?

- Leg de gehanteerde analysemethode uit en meet de nauwkeurigheid.
  - Is in de technische documentatie beschreven welke methode(s) is/zijn gebruikt?
  - Is de gehanteerde analysemethode uitgelegd?
  - Is de nauwkeurigheid van de analysemethode gemeten en beschreven?

- Leg de inputgegevens (brondata/datasets) vast die gebruikt worden en gebruik daarbij enkel relevante data. Documenteer.
  - Is in de technische documentatie vastgelegd welke databronnen zijn gebruikt?
  - Zijn de inputgegevens (brondata/datasets) die gebruikt worden vastgelegd?
  - Zijn de gebruikte data relevant?
  - Zijn de gebruikte data en de relevantie ervan gedocumenteerd?

- Beschrijf de kwaliteit van de gebruikte databron(nen) en of deze kwaliteit voldoende is voor het doel waarvoor deze wordt ingezet.
  - Is in een '*data deep dive*' de kwaliteit van de databronnen grondig onderzocht en toegelicht? Is ook in de technische documentatie beschreven hoe de datakwaliteit wordt onderzocht? Worden alleen databronnen met voldoende kwaliteit meegenomen in het model?
  - Is de kwaliteit van de gebruikte databron(nen) onderzocht?
  - Is de kwaliteit van de gebruikte databron(nen) beschreven?

- Leg de aannames/keuzes die gehanteerd zijn vast. Denk aan keuzes zoals dat bepaalde data niet is meegenomen in de analyse, omdat de kwaliteit daarvan onvoldoende is of omdat een dataset onvoldoende gegevens bevat om een statistische analyse op uit te voeren.
  - Zijn alle keuzes beschreven in de technische documentatie?
  - Zijn alle keuzes gedocumenteerd?

- Collegiale uitlegbaarheid: zorg ervoor dat teams volledig toegang/inzicht hebben in elkaars documentatie, beslissingen en code. Wanneer beslissingen over *features*, specificaties, ontwerp, bouw en tests verdeeld zijn over meerdere teams kunnen er in de overdracht ongemerkt en onbedoeld interpretatie-verschillen ontstaan. Transparantie en uitlegbaarheid komen dan in gevaar.
  - Hoe is collegiale uitlegbaarheid gewaarborgd?

Gegevensherkenning

Wijze van parametrisering, keuze voor trainingsgegevens en verkenning van de potentiële discriminerende factoren. Vastlegging van keuzes voor reproduceerbaarheid. *Sensitivity analysis* van keuzes van parametrisering, inzet trainingsdata.

- Is de wijze waarop parametrisering tot stand is gekomen beschreven?
- Is de wijze waarop de keuze voor trainingsgegevens tot stand is gekomen beschreven?
- Zijn potentiële discriminerende factoren verkend en gedocumenteerd?
- Is vastgelegd welke trainingsgegevens gebruikt zijn?
- Is vastgelegd welke parameters gebruikt zijn?

- Is geanalyseerd en beschreven wat de gevolgen zijn van een andere parameterkeuze voor de resultaten?
- Is geanalyseerd en beschreven wat de gevolgen zijn van een andere trainingsdataset voor de resultaten?
- Is al het bovenstaande beschreven in de technische documentatie?

- Is er onderscheid gemaakt tussen de gegevens de volgende drie soorten gegevens?
  - De gegevens die worden gebruikt om een model te trainen of te ontwikkelen.
  - De gegevens die als onderdeel van het proces worden verzameld/ingevoerd.
  - De gegevens die worden opgevraagd uit andere processen of anderen bronnen.

- Aanvulling vanuit [18]:
  - Hoe wordt er omgegaan met bias in de data?
  - Hoe wordt er omgegaan met fouten in de data?
  - Hoe wordt er omgegaan met outliers in de data?
  - Hoe wordt er omgegaan met missende waarden?

Auditeerbaarheid

Focus op proces en vastlegging van proces voor verificatie. Dit betekent een gedegen R&D-proces plus documentatie waarin het gebruik van algoritmen in productie navolgbaar is.

Met als specifieke aandachtspunten en maatregelen:

- Werk niet met confidentiële algoritmen, maar met open algoritmen, die toegankelijk zijn voor controllers en toezichthouders en bij voorkeur ook voor experts en burgers. Dat impliceert het volgende:
  - Het algoritme dient niet-confidentieel te zijn;
    - Is het algoritme niet-confidentieel?
    - Wordt het algoritme niet gepubliceerd, maar mag het wel ge-audit worden?
  - Het algoritme dient gedocumenteerd te zijn;
    - Is bij de code beschreven wat er gebeurt?
    - Is een technische documentatie beschikbaar waarin de modelontwikkeling is beschreven?
  - Gebruik zoveel mogelijk algoritmen en analysemethoden die wetenschappelijk gevalideerd zijn;
    - Zijn er uitsluitend algoritmen gebruikt die wetenschappelijk gevalideerd zijn?
    - Zijn de algoritmen/analysemethoden (zoveel mogelijk) wetenschappelijk gevalideerd?
  - Gebruik indien mogelijk algoritmen die reeds *open source* zijn of stel deze als *open source* beschikbaar.
    - Is het algoritme *open source* (indien mogelijk)?
    - Zijn er alleen *open source* technieken gebruikt (zoals *Random Forest*)?
    - Is het algoritme gepubliceerd?

  Er kunnen redenen zijn om van de bovenstaande richtlijnen op basis van een juiste onderbouwing af te wijken.

- Onderbouwing van keuzes.
  - Is in de technische documentatie beschreven waarom het uiteindelijke algoritme gekozen is en welke data gebruikt is?
  - Is de keuze voor de gebruikte algoritmen onderbouwd?

  - Is de keuze voor de gebruikte data onderbouwd?

- Noteer waarnemingen, zoals afwijkingen in de gegevens of onverwachte/onverklaarbare resultaten.
  - Is na elke oplevering beschreven of er veranderingen zijn in de gegevens en of er afwijkende resultaten zijn?
  - Zijn waarnemingen zoals afwijkingen in gegevens of onverwachte/ onverklaarbare resultaten genoteerd?

- Gebruik eenvoudige methoden boven complexe methoden daar waar mogelijk. Dit komt ten goede aan de uitlegbaarheid, auditeerbaarheid en beperking van risico's.
  - Heeft er een vergelijking plaatsgevonden tussen simpelere en complexere algoritmen?
  - Wat voor tussenvormen hebben jullie bedacht tussen oude en nieuwe situatie?
  - Is voor de meest simpele methode gekozen die toch het doel van het algoritme bereikt?
  - Is de keuze voor kunstmatige-intelligentie-technieken gekozen, omdat dit tot veel betere resultaten leidt dan een simpele methode?
  - Behaalt het algoritme zonder de complexe methode niet het doel waarvoor het ontwikkeld is?

- Vergelijking complexere en simpelere methoden op specifieke input.
  - Is validatie onderdeel van de modelontwikkeling waarin een aantal verschillende methoden zijn gebruikt en geverifieerd?
  - Zijn de uitkomsten van het algoritme logisch, gegeven de data-input? (voor mensen logisch)
    - Hoe is dit geëvalueerd?

- Lever een gedetailleerde omschrijving van het algoritme en de werking ervan, samen met een controleerbare validatie dat de code overeenkomt met de specificatie.
  - Is in de technische documentatie is de werking van het algoritme gedetailleerd beschreven?

- Zorg voor reproduceerbaarheid.
  - Kan men de code van het model hergebruiken en precies nadoen hoe de eerste modelresultaten waren, mits de data uit die tijd kunnen en mogen worden gebruikt?
  - Zijn alle wijzigingen aan de code bijgehouden, waardoor het mogelijk is om terug te gaan naar de eerste versie van het model?
  - Zijn de resultaten van het algoritme reproduceerbaar?

- Zorg dat er bij algoritmen die in besluitvorming uitmonden dan wel anderszins een aanmerkelijke impact op burgers, bedrijven of samenleving, een vorm van betekenisvolle menselijke tussenkomst is georganiseerd.[3]
  - Hoe is menselijke tussenkomst georganiseerd?

- Aanvulling vanuit EC: Menselijk toezicht.
  - Hebt u nagedacht over de gepaste mate van menselijke controle voor dit specifieke AI-systeem en deze specifieke gebruikssituatie?
  - Kunt u, indien van toepassing, de mate van menselijke controle of betrokkenheid beschrijven? Wie is de "*human in control*" en wat zijn de momenten of hulpmiddelen voor menselijke interventie?
  - Hebt u mechanismen en maatregelen ingesteld om dergelijke potentiële menselijke controle of menselijk toezicht te waarborgen of om te zorgen

dat beslissingen onder de algehele verantwoordelijkheid van mensen worden genomen?
  - Hebt u maatregelen genomen om controle mogelijk te maken en problemen in verband met het beheren van AI-autonomie te verhelpen?
  - Indien het AI-systeem of de gebruikssituatie zelflerend of autonoom is: hebt u specifiekere controle- en toezichtsmechanismen ingesteld?
  - Wat voor detectie- en responsmechanismen hebt u ingesteld om te controleren of er iets mis kan gaan?

- Aanvulling vanuit [18]:
  - Is er een *baseline*-model ontwikkeld dat intrinsiek uitlegbaar is? Denk hierbij aan lineaire regressie, logistische regressie en *decision tree* modellen.
  - Is er een vergelijking uitgevoerd tussen complexere modellen en het baseline model?
  - Zijn er specifieke methodes gebruikt om de uitlegbaarheid van complexe modellen te vergroten?

- Aanvulling vanuit [18]:
  Tevens is hier van belang het documenteren/annoteren in de code wat een routine doet of moet doen om een goede code-review uit te kunnen laten voeren door b.v. een andere analist.
  - Zijn alle stappen in de analyse gecodeerd? Is ervoor gezorgd dat er geen handmatige acties nodig zijn?
  - Is er gezorgd voor versiebeheer op de code?
  - Is geanonimiseerde trainingsdata bewaard?

Verantwoording

Overheden zijn verantwoordelijk voor de ontwikkeling en inzet van hun algoritmen en dienen daarover dan ook verantwoording af te leggen.

- Wanneer meerdere organisaties betrokken zijn bij de ontwikkeling van algoritmische systemen, wordt geregeld dat één organisatie het voortouw en regie heeft bij de ontwikkeling van het systeem en als zodanig primair voor de toepassing van de richtlijnen en de uitvoering van andere relevante instrumenten, zoals DPIA.[4]
  - Zijn er meerdere organisaties betrokken bij ontwikkeling? En zo ja, wie heeft het voortouw m.b.t. verantwoordelijkheid en uitvoering van b.v. de DPIA?

- Bij complexe algoritmische systemen dient binnen de eigen organisatie de regie en bijbehorende verantwoordelijkheid voor het ontwerp- en ontwikkelproces duidelijk te worden belegd en worden geborgd dat ook na implementatie van het systeem, de regie en verantwoordelijkheid helder belegd is.
  - Hoe is de verantwoordelijkheid van ontwerp en ontwikkelproces belegd?
  - Zijn ook de regie en verantwoordelijkheid helder belegd na implementatie van het systeem?

- Afwijking van richtlijn van verantwoording. De verantwoording gebeurt volgens het principe van *comply or explain*. Dit wil zeggen dat als uitgangspunt geldt dat organisaties de richtlijn moet volgen, of nadrukkelijk uitleggen waarom zij ervan afwijken. Dit geldt in eerste instantie voor algoritmen die nieuw worden ontwikkeld, maar ook voor algoritmen die reeds werkzaam zijn. (Gelet daarop wordt de toepassing van de richtlijnen bij voorkeur verankerd in het kader van een P&C-cyclus, waarbij jaarlijks door organisaties wordt gerapporteerd over situaties waarin besloten is om af te wijken van de richtlijnen.)
  - Wordt er afgeweken van hoe verantwoording is aangegeven in de richtlijn?

    - Zo ja: *(Comply or) explain*?
    - Zo ja: Wordt er jaarlijks gerapporteerd hoe er wordt afgeweken van de richtlijn?

- Zorg dat de voor verantwoording noodzakelijke informatie (zoals documentatie, broncode of data) beheerd wordt conform de eisen die de Archiefwet stelt.
  - Wordt voor de verantwoording noodzakelijke informatie (documentatie, broncode of data) conform de eisen van het Archiefwet beheerd?

Validatie

Nadruk op gedocumenteerde validatie van het model in het perspectief van het beoogde doel.

- Gebruikt de overheidsinstantie-strikte methoden om het model te valideren?
- Documenteert de overheidsinstantie de (validatie)methode en de resultaten?
- Test en beoordeelt de overheid routinematig of het model het beoogde doel bereikt en geen bijkomende schade oplevert?
- Indien mogelijk, publiceert de overheidsinstantie deze testresultaten?

Aanvulling vanuit [18]:

- Worden de modellen met een grote potentiële impact gevalideerd door een onafhankelijke partij?
- Wordt het model continu gemonitord?
- Uitgangspunt is dat de overheidsinstantie over deze testresultaten publiceert. Wordt dit gedaan?

Toetsbaarheid

Toetsbaarheid betreft de inrichting van methode van data-analyse, de gehanteerde algoritmen, datasets en de feitelijke verwerkingen, zodat deze daadwerkelijk kunnen worden getoetst. Dit richt zich primair op toetsing door de toezichthouder en de rechter.

- De bestuursrechter stelt op grond van het bestuursrecht eisen aan de inzichtelijkheid, controleerbaarheid en toegankelijkheid in geval van geautomatiseerde besluitvorming door bestuursorganen.
  Voor het toetsingskader voor de beoordeling van geautomatiseerde besluitvormingsprocessen m.b.v. algoritmen, zie de uitspraak van de Afdeling bestuursrecht van de Raad van State van 17 mei 2017 en voornoemd ongevraagd advies van de Raad van State van 31 augustus 2018.
- Er zijn op het moment door MinJenV geen concrete punten noch vragen gemeld bij deze richtlijn. Om deze reden zijn er in de technisch-inhoudelijke evaluatie geen directe vragen gesteld met betrekking tot toetsbaarheid.

Publieksvoorlichting

Het betreft hier de algemene informatie voor het publiek en niet de individuele voorlichting, waarin bijvoorbeeld informatie over de persoonsgegevens worden verwerkt. In plaats van interviewvragen is de richtlijn besproken aan de hand van voorwaarden aan publieksvoorlichting en onderwerpen waarover een overheidsdienst publiek dient te informeren.

Voorwaarden van publieksvoorlichting:

- Beknopt en transparant (efficiënt en bondig, gelaagde informatie).
- Begrijpelijk (gemiddelde vertegenwoordiger van beoogd publiek).
- Gemakkelijk toegankelijk (met weinig moeite).

Als een overheidsdienst data-analyses verricht, dient deze dienst op haar website het publiek te informeren over:

- dat zij data-analyses uitvoert;
- waarom zij data-analyses uitvoert (wat het doel ervan is en wat met de resultaten daarvan wordt gedaan);
- waarom het gebruik van data-analyses proportioneel is, en er geen betere alternatieven waren om het doel te bereiken;
- wat de eventuele consequenties van de analyse voor betrokken burgers zijn, en hoe rekening is gehouden met eventuele impacts ervan op grondrechten;
- eventuele toepassing van *machine learning* en de uitleg daarvan;
- wat de wettelijke grondslag voor het uitvoeren van deze analyses is,
- welke databronnen van welke organisaties daarvoor worden gebruikt, en wat de kwaliteit daarvan is;
- welke persoon binnen de overheidsorganisatie verantwoordelijk voor de analyse is;
- wat de rol van eventuele derden bij deze analyses is;
- met welke organisaties (publiek of privaat), indien dit aan de orde is, brondata en/of resultaten van de analyses worden gedeeld;
- welke kwaliteitsborging er plaatsvindt (welke risico's worden onderkend en welke maatregelen daartegen worden genomen en op welke wijze toetsing plaatsvindt);
- hoe er tussen analyse en een eventueel besluit menselijke tussenkomst plaatsvindt,
- welke toetsingskaders er zijn.

Bijlage 2: Resultaten van de interviews per richtlijn

Deze bijlage bevat de resultaten van de interviews die TNO heeft gehouden met vertegenwoordigers van zowel gemeente Nissewaard als Totta Data Lab. Hierbij zijn qua structuur de richtlijnen van MinJenV als basis genomen, die ook de basis zijn van de vragenlijst zoals te vinden in Bijlage 1. Voor de definities van de richtlijnen refereren wij naar de richtlijnen van MinJenV [7, 18]. Voor toetsbaarheid zijn geen vragen aanwezig, dus deze is niet meegenomen in de interviews.

Bewustzijn en inperking van risico's
Vanuit Nissewaard is in kaart gebracht wat de risico's zijn van het gebruik van het AI-algoritme. Er zijn ook maatregelen door Nissewaard getroffen om fouten en ongewenste bias te voorkomen. Eén van de overwegingen is welke variabelen wel en welke niet meegenomen dienen te worden in het gebruikte AI-algoritme. Zo hebben de variabelen leeftijd en geslacht wel een toegevoegde waarde, terwijl deze variabelen bij b.v. een sollicitatie zouden kunnen zorgen voor bias. Hierbij laat Nissewaard bewustzijn van risico's zien, doordat leeftijd en geslacht enkel zijn meegenomen nadat duidelijk werd dat deze toegevoegde waarde hadden voor de nauwkeurigheid van het algoritme. Eén van de risicobeperkingen is dat bij de implementatie van het algoritme verschillende partijen zijn meegenomen, zoals vooronderzoek bij andere gemeenten waar al een algoritme wordt gebruikt, het gebruiken van de landelijke klankbordgroep (die ook verbonden zijn met *een transparency lab)*, een gemeentelijke klankbordgroep en een gemeentelijke advies-/cliëntenraad. Daarnaast zijn de Functionaris Gegevensbescherming en Juridische Zaken, sociale rechercheurs en teamleiders meegenomen in het proces. Ook in de evaluatie van Nissewaard worden de risico's en maatregelen geëvalueerd, bijvoorbeeld door de kwaliteit van de variabelen te controleren en een jaarlijkse opfriscursus over het AI-algoritme voor de rechercheurs. Een genoemd risico in deze evaluatie van Nissewaard is dat het kan voorkomen dat mogelijke fraudes als negatief gelabeld worden als er onvoldoende bewijs gevonden is. De gemeente kan niet veel anders. Dit heeft echter wel impact op het systeem, aangezien daarop verder getraind wordt en hiermee de (nauwkeurigheid van de) evaluatie lastig maakt. Een belangrijk aspect is dat de risico's verminderd zijn doordat het algoritme momenteel maar 10-15% van het werk uitmaakt voor de rechercheurs, en doordat het merendeel van het werk vanuit passieve signalen zoals meldingen van bijvoorbeeld een buurman komt. Het algoritme is dus een aanvullend instrument om signalen van misbruik en oneigenlijk gebruik te genereren. Het vervangt hier de zogenaamde algemene hercontroles, waarin jaarlijks de rechtmatigheid van bijna iedere uitkering aan onderzoek werd onderworpen.

Vanuit TDL zijn ook de risico's en maatregelen overwogen, waarbij er inhoudelijk ook nadrukkelijk wordt onderzocht wat de biases en fouten zijn. Bij de start van elk onderzoek wordt een '*data deep dive*' gedaan, waarin er wordt geanalyseerd of de kwaliteit en kwantiteit van de data voldoende zijn om een model erop te trainen. Hiertoe wordt op macro- en microniveau onderzoek gedaan naar verandering in de aanlevering van de data, koppelbaarheid, variatie, vulling, en bias (voor elke variabele). Afwijkingen worden besproken met Nissewaard. Verder wordt er na het trainen van een nieuw model bij elke oplevering onderzocht of het model op eenzelfde manier heeft geleerd middels een inspectie van de regels en de top-10 of top-100. Ook bij de ontwikkeling van het model is er rekening gehouden met de risico's van het algoritme. Er zijn verschillende algoritmen onderzocht, en is gekozen voor een ensemble (combinatie) om zo de kracht van deze algoritmen te bundelen, en zo het risico op fouten te verminderen. Er wordt in het proces rekening gehouden met transparantie, wat de keuze voor dit ensemble van algoritmen ook heeft beïnvloed. Doordat het model ook gevoed wordt door meldingen en door menselijke validatie, is het minder vatbaar om in een

ongewenste lus te komen, waarin het leert van zichzelf en daardoor mogelijke biases kan versterken.

Uitlegbaarheid
Er is veel gedocumenteerd t.a.v. het gebruikte AI-algoritme, de gemaakte keuzes tussen (bijvoorbeeld) analysetechnieken en alle 40 gebruikte variabelen. In de technische documentatie zijn de databronnen, relevante inputdata, analyse van de methode en performance (nauwkeurigheid) beschreven. De uitleg van de '*data deep dive*', de onderbouwing van de relevantie van de gekozen variabelen en de precieze beslisregels voor het algoritme zijn hierin niet, of niet volledig, beschreven; dit is ook niet wettelijk verplicht.

Voor de collegiale uitlegbaarheid is vanuit TDL de communicatie met Nissewaard middels presentaties met handhavers geregeld. Tijdens deze presentaties worden verscheidene soorten data die gebruikt worden besproken en ook waarom het belangrijk is om deze goed te registeren. Intern wordt de uitlegbaarheid gedaan via GitHub en de technische documentatie. In de evaluatiesessies wordt besproken welke veranderingen zijn doorgevoerd.

Gegevensherkenning
Zowel de keuze voor de trainingsgegevens als de parametrisering is door TDL overwogen. De keuze voor de trainingsgegevens wordt in overleg tussen Nissewaard en TDL bepaald aan de hand van de *data deep dive*. Hierbij worden potentiële discriminerende factoren gesignaleerd zoals de variabele "bijzonder situaties" (b.v. "heeft geen woonhuis"). Deze worden alleen meegenomen als er voldoende voorbeelden hiervan zijn in de data. Ook kunnen alle categorische variabelen discriminerend zijn, maar daartoe wordt door TDL een bias-analyse in de *data deep dive* gedaan. In de keuze van de parameters is er rekening gehouden met de *trade-off* tussen performance (nauwkeurigheid) en begrijpelijkheid. Dit wordt per gemeente opnieuw bekeken. Na de keuze voor de trainingsgegevens wordt er een model getraind, waarbij de parameters bestaan uit de diepte van de boom, aantal bomen en de zogeheten lambda -parameter in het *rule-fit* model [16]. In de modelevaluatie wordt er een cross-validatie gedaan als controlemechanisme of het model robuust is.

Auditeerbaarheid
Het gebruik van databronnen voor training en testen zijn gedocumenteerd. Het algoritme is gebaseerd op het "*rule-fit*" model welke openbaar beschikbaar is. De specifieke implementatie en gemaakte keuzen zijn niet openbaar beschikbaar, maar zijn wel auditeerbaar.

Er is specifiek gekozen voor een uitlegbaar en transparant algoritme, dat tevens reproduceerbaar is. Dit is ook beschreven in de technische documentatie. Per oplevering wordt het model gecontroleerd, en waar nodig wordt het algoritme verbeterd. De verandering wordt vastgelegd, en ook onverwachte resultaten worden vastgelegd en besproken in een evaluatiesessie. Het algoritme wordt alleen gebruikt als een suggestie, en de top-10 resultaten worden onderzocht door rechercheurs van Nissewaard. De gehele ranking wordt wel aan Nissewaard geleverd. Er is niet gekozen voor tussenvormen van aanvullend onderzoek van de top-50 of top-100 wegens beperkte capaciteit.

Verantwoording
In de gemeente zijn de volgende functies betrokken bij het proces: privacy officer, chief information security officer (ciso), functionaris gegevensbeschermer, manager (ambtelijk verantwoordelijk), wethouder, ketenregisseur, applicatiebeheerder, beleidsadviseur, sociaal rechercheur, en in de beginfase Stimulansz (dit is de juridische invalshoek op sociaal domein). De hoofd- / eindverantwoordelijke is de wethouder. Er is een DPIA opgesteld door Nissewaard (en publiekelijk opengesteld)

en de archiefwet wordt nageleefd. Er is tevens een DEDA (De Ethische Data Assistent) afgelegd via de Utrecht Data School.

Validatie
Het einddoel van het systeem van Nissewaard is om de nalevingsbereidheid zo hoog mogelijk te krijgen, dus niet om zoveel mogelijk mensen te pakken op misbruik en oneigenlijk gebruik van de bijstand. In dit proces is het belangrijk dat de inwoners en de rechercheurs zo min mogelijk belast worden. Er wordt niet alleen gekeken naar passieve signalen (meldingen), maar ook de actieve signalen (thema-controles en het algoritme). Hoe goed het einddoel wordt behaald is lastig te meten, maar kan door proxy's zoals gemiddeld opgespoord fraudebedrag wel indirect worden geschat. De resultaten worden gedocumenteerd en er zijn periodieke checks (opleveringen). Daarnaast zijn de besproken vragen en antwoorden van de raad openbaar.

Publieksvoorlichting
Nissewaard heeft veel moeite gestopt in publieksvoorlichting. Er is gesproken met het merendeel van de grote kranten, er zijn vragen uit de raad beantwoord en alle vragen van geïnteresseerden zijn beantwoord. Daarnaast zijn er verschillende documenten openbaar beschikbaar gemaakt, variërend van een infographic tot DPIA tot technische documentatie.