**Nota** voor burgemeester en wethouders
**Onderwerp**
Evaluatie Pilot data analyse Inkomensondersteuning

Team
DEV-ASK

| 1- Notagegevens | |
|---|---|
| Notanummer | 2021-000527 |
| Datum | 08-03-2021 |
| Programma: | |
| 07 Inkomensvoorziening en arbeidsmarkt | |
| Portefeuillehouder Weth. De Geest | |

| 2- Bestuursorgaan | |
|---|---|
| [X]B & W | 16-03-2021 |
| [ ]Raad | -- |
| [ ]Burgemeester | -- |

| College van B & W | |
|---|---|
| - Burgemeester | - Weth. Grijsen |
| - Weth. De Geest | - Weth. Verhaar |
| - Weth. Walder | - Weth. Rorink |

| Besluitenlijst | d.d. | | d.d. | | d.d. |
|---|---|---|---|---|---|
| [ ]Akkoordstukken | -- | [X]Openbaar | 16-03-2021 | [ ]Besloten | -- |

| Routing | d.d. |
|---|---|
| Burgemeester | 09-03-2021 |
| Wethouder | 10-03-2021 |
| Programmamanager | 10-03-2021 |
| Programmamanager | 09-03-2021 |

| | par. |
|---|---|
| [ ]adj.secr. | -- |
| [X]gem.secr. | 10-03-2021 |
| BIS Openbaar | |
| Status | Definitief2021-03-17 |

Bijlagen
Evaluatierapport Pilot geautomatiseerde Data analyse bijstandsuitkering
Advies Pilot bijstand
Technische documentatie algoritme Deventer
Onderzoeksprotocol Sociale Recherche pilot data analyse

B & W d.d.: 16-03-2021

Besloten wordt:
1 Kennis te nemen van het evaluatierapport Pilot Geautomatiseerde data analyse bijstandsuitkering;
2 de conclusie en het advies uit het evaluatierapport over te nemen en voor het vervolg te betrekken bij het project Datagedreven Werken;
3 de raadsmededeling vast te stellen en aan te bieden aan de raad;
4 de nota en het besluit openbaar te maken.

| **Financiële aspecten:** | |
|---|---|
| Financiële gevolgen voor de gemeente? | Nee |
| Begrotingswijziging | Nee |

**Voorstel openbaarmaking conform Wet Openbaarheid Bestuur (Wob)**

[X] De nota en het besluit openbaar te maken
[ ] De nota en het besluit openbaar te maken vergezeld van bijgaand persbericht
[ ] De nota en het besluit openbaar te maken nadat

[ ] De nota en het besluit openbaar te maken, behalve…

[ ] Het besluit openbaar te maken, maar niet de nota, gelet op artikel:

[ ] De nota en het besluit niet openbaar te maken, gelet op artikel:

gemeente Deventer

**Kennisgeving/ Bekendmaking Awb**

| | |
|---|---|
| Kennisgeving (publicatie) conform Awb | Nee |
| Bekendmaking conform Awb | Nee |

**ADVIESRADEN:**

| | |
|---|---|
| Moet een van de adviesraden gehoord worden of op de hoogte gesteld? | Nee |

# Toelichting

## Inleiding

In de periode juli 2018 - september 2019 heeft de gemeente een pilot data analyse met behulp van een zelflerend algoritme uitgevoerd. De pilot werd uitgevoerd in samenwerking met het bedrijf Totta Data Lab uit Amsterdam en richtte zich op het opsporen van fraude door middel van kansberekeningen gemaakt door een zelflerend algoritme.

De pilot was als onderzoek opgezet met daarbij behorende onderzoeksvragen.
Centrale onderzoeksvraag:

**Leidt geautomatiseerde data-analyse tot betere controle van bijstand?**
De centrale onderzoekvraag is in volgende deelvragen onderverdeeld:

- Zijn we door middel van data-analyse gestuurd werken in staat om effectiviteit van onze onderzoeken te verhogen?
- Zijn we door middel van data-analyse gestuurd werken in staat onze dienstverlening te verbeteren?
- Kan data-analyse gestuurd werken uitgebreid worden naar andere werkgebieden (bijvoorbeeld Wmo, schuldhulpverlening)?

**Kwantitative resultaten**

| Ronde | Aantal hoogst scorende | In onderzoek genomen | Uitkering beëindigd | Score | Bijzonderheden |
|---|---|---|---|---|---|
| 1 | 10 | 8 | 3 | 30% | Naast 3 beëindigingen zijn twee dossiers door DWT opgepakt om traject richting arbeid weer te starten. |
| 2 | 10 | 3 | 0 | 0% | |
| 3 | 10 | 2 | 1 | 10% | En drie maanden uitkering teruggevorderd. |
| 4 | 10 | 8 | 0 | 0% | |
| Totaal | 40 | 21 | 4 | 10% | |

**Conclusie van de pilot**
Leidt geautomatiseerde data-analyse tot betere controle van bijstand?
Ja. Het rendement van soortgelijke onderzoeken (themacontroles) kwam doorgaans op 3% uit, rendement pilot 10%. Daarbij moet worden meegenomen dat deze manier van controle aanvullend moet zijn aan andere methoden en technieken w.o. fraudealertheid van medewerkers. Daarin verandert niets. Zich baseren op één controlemethode was en blijft gevaarlijk.

Daarnaast kan worden geconcludeerd dat:

- Er binnen de Participatiewet voldoende juridische kaders zijn waarbinnen het werken met algoritmen mogelijk is.
- Deze manier van controle van bijstand loont. De kosten zijn lager dan de baten. Het is natuurlijk niet altijd inzichtelijk te maken welke baten er zijn en welke financiële waarde ze vertegenwoordigen,

maar inzetten van algoritmen kan in de basis tegen beperkte kosten.

**Advies**

1. Maak werken met zelflerende algoritmen vast onderdeel van het signaalgestuurd werken binnen de Participatiewet.
2. Onderzoek verder door pilots/projecten uit te voeren met verschillende onderwerpen en wettelijke kaders.
3. Sluit aan bij bestaande ontwikkelingen binnen de gemeente Deventer zoals Datagedreven Werken, Deltaview. e.d.
4. Trek samen op met andere gemeenten, wetenschap en landelijke organisaties.

## Beoogd resultaat

Het ontwikkelen van handhavingsmogelijkheden ten behoeve van rechtmatige verstrekking van de bijstandsuitkering.

Uit het regeerakkoord:
“Misbruik van sociale voorzieningen ondermijnt het draagvlak voor solidariteit. Het kabinet vindt het van belang dat uitvoerders, waaronder gemeenten, effectief gebruik maken van de mogelijkheden tot het delen, koppelen en analyseren van data, uiteraard met inachtneming van de geldende wettelijke regels en waarborgen. Dit kan ook uitkeringsgerechtigden helpen om regels na te leven en fouten te voorkomen.”

## Kader

1. Participatiewet
2. AVG
3. Beleidsplan “Dienstverlenend Handhaven”, handhavingsbeleid 2018-2022
4. Juridisch advies van Eiffel Legal Office

## Argumenten voor en tegen

Voor

- Zelflerende algoritmen kunnen als hulpmiddel gemeentelijke dienstverlening efficiënter en effectiever maken.
- Ontwikkelen van algoritmen kan met beperkte financiële middelen.
- Er zijn inmiddels verschillende hulpmiddelen en richtlijnen ontwikkeld waar de gemeente gebruik van kan maken om algoritmen juridisch en ethisch verantwoord in te zetten.

Tegen
Zoals in het evaluatierapport wordt aangegeven is het inzetten van zelflerende algoritmen niet zonder risico. De zogenaamde BIAS (onbedoeld discrimineren) ligt op de loer. Het afleggen van verantwoording en transparantie van besluitvorming kunnen in geding komen.

## Extern draagvlak (partners)

Het werken met zelflerende algoritmen is nieuw. Het is van belang om samenwerking te blijven zoeken met cliëntenraden, VNG en wetenschappelijke instanties.

## Financiële consequenties

Het financieel resultaat van de pilot:

| | |
|---|---|
| Opbrengsten | € 56.800 |
| Kosten | € 41.500 |
| Resultaat | € 15.300 |

Toekomstige financiële gevolgen zullen per pilot/project in kaart moeten worden gebracht.

## Aanpak/uitvoering

1. Bij het organiseren van een volgende pilot/project gelden de volgende kritische succesfactoren:

   - Het juridisch kader is van tevoren vastgesteld.
   - Ethische overwegingen en kaders zijn van tevoren bekend (DEDA).
   - Formuleer een onderzoeksvraag. Deze onderzoeksvraag richt zich niet alleen op het inhoudelijke onderwerp, maar ook op andere randvoorwaardelijke zaken. Zoals: is onze data van goede kwaliteit, hebben we de juiste competenties om met zelflerende algoritmen te werken, etc.
   - Zorg voor een multidisciplinaire teams, waar inhoudelijk betrokken collega's worden aangevuld met collega's van ICT, Kennis en Verkenning, Privacy.
   - Betrek project Datagedreven werken bij de volgende pilots/projecten. Op deze manier vergroot de gemeente het lerend vermogen van haar organisatie. Alle kennis die in pilots/projecten wordt opgedaan, kan zo gebundeld worden en kan de gemeente als organisatie de stap naar datagestuurd werken zetten. Deel na afronding van een pilot via de data academie opgedane kennis met de hele organisatie, communiceer openlijk.
   - Regel bestuurlijk commitment bij elke pilot/project.
   - Werk toe naar een centrale plek waaruit werken met zelflerende algoritmen en andere vormen van kunstmatige intelligentie aangestuurd en begeleid wordt.

gemeente Deventer

# RAADSMEDEDELING

**Onderwerp** Evaluatie Pilot data analyse Inkomensondersteuning

| | | | |
|---|---|---|---|
| **Mededelingennr** | 2021-000527 | **Portef.houder** | Weth. De Geest |
| **Team** | DEV-ASK | **BenW-besluit d.d.:** | 16 maart 2021 |

## 1. Inleiding: waarom deze mededeling

In de periode juli 2018 - september 2019 heeft de gemeente de pilot data analyse bijstandsuitkeringen uitgevoerd met behulp van een zelflerend algoritme. De pilot werd uitgevoerd in samenwerking met het bedrijf Totta Data Lab uit Amsterdam en richtte zich op het opsporen van fraude met behulp van kansberekeningen gemaakt door een zelflerend algoritme. Inmiddels is het evaluatierapport van de pilot beschikbaar.

## 2. Kader

- Participatiewet
- AVG
- Beleidsplan “Dienstverlenend Handhaven”, handhavingsbeleid 2018-2022
- Juridisch advies van Eiffel Legal Office

## 3. Kern van de boodschap

De conclusie van de uitgevoerde pilot en het evaluatierapport is dat geautomatiseerde data-analyse leidt tot een betere controle van de bijstand, mits uitgevoerd onder de juiste juridische en ethische randvoorwaarden en als aanvulling op andere methodes en technieken. Dit bevordert de rechtmatige verstrekking van de bijstandsuitkering. De opbrengst van deze pilot wordt meegenomen bij het project Datagedreven Werken waarover de raad binnenkort verder geïnformeerd wordt.

## 4. Nadere toelichting

De pilot was als onderzoek opgezet met daarbij behorende onderzoeksvragen.

**Centrale onderzoeksvraag:**
Leidt geautomatiseerde data-analyse tot betere controle van bijstand?

De centrale onderzoekvraag is in volgende deelvragen onderverdeeld:

- Zijn we door middel van data-analyse gestuurd werken in staat om effectiviteit van onze onderzoeken te verhogen?
- Zijn we door middel van data-analyse gestuurd werken in staat onze dienstverlening te verbeteren?
- Kan data-analyse gestuurd werken uitgebreid worden naar andere werkgebieden (bijvoorbeeld Wmo, schuldhulpverlening)?

**Kwantitative resultaten**

| Ronde | Aantal hoogst scorende | In onderzoek genomen | Uitkering beëindigd | Score | Bijzonderheden |
|---|---|---|---|---|---|
| 1 | 10 | 8 | 3 | 30% | Naast 3 beëindigingen zijn twee dossiers door DWT opgepakt om traject richting arbeid weer te starten. |
| 2 | 10 | 3 | 0 | 0% | |
| 3 | 10 | 2 | 1 | 10% | En drie maanden uitkering teruggevorderd. |
| 4 | 10 | 8 | 0 | 0% | |
| Totaal | 40 | 21 | 4 | 10% | |

**Conclusie van de pilot**
Leidt geautomatiseerde data-analyse tot betere controle van bijstand?

Ja. Het rendement van soortgelijke onderzoeken (themacontroles) kwam doorgaans op 3% uit, rendement pilot 10%. Daarbij moet worden meegenomen dat deze manier van controle aanvullend moet zijn aan andere methoden en technieken w.o. fraudealertheid van medewerkers. Daarin verandert niets. Zich baseren op één controlemethode was en blijft gevaarlijk.

Daarnaast kan worden geconcludeerd dat:

- Er binnen de Participatiewet voldoende juridische kaders zijn waarbinnen het werken met algoritmen mogelijk is.
- Deze manier van controle van bijstand loont. De kosten zijn lager dan de baten. Het is natuurlijk niet altijd inzichtelijk te maken welke baten er zijn en welke financiële waarde ze vertegenwoordigen, maar inzetten van algoritmen kan in de basis tegen beperkte kosten.

**Advies**

1. Maak werken met zelflerende algoritmen vast onderdeel van het signaalgestuurd werken binnen de Participatiewet.
2. Onderzoek verder door pilots/projecten uit te voeren met verschillende onderwerpen en wettelijke kaders.
3. Sluit aan bij bestaande ontwikkelingen binnen de gemeente Deventer zoals Datagedreven Werken, Deltaview. e.d.
4. Trek samen op met andere gemeenten, wetenschap en landelijke organisaties.

Bij het organiseren van een volgende pilot/project gelden de volgende kritische succesfactoren:

- Het juridisch kader is van tevoren vastgesteld.
- Ethische overwegingen en kaders zijn van tevoren bekend (DEDA).
- Formuleer een onderzoeksvraag. Deze onderzoeksvraag richt zich niet alleen op het inhoudelijke onderwerp, maar ook op andere randvoorwaardelijke zaken. Zoals: is onze data van goede kwaliteit, hebben we de juiste competenties om met zelflerende algoritmen te werken, etc.
- Zorg voor een multidisciplinaire teams, waar inhoudelijk betrokken collega's worden aangevuld met collega's van ICT, Kennis en Verkenning, Privacy.
- Betrek project Datagedreven werken bij de volgende pilots/projecten. Op deze manier vergroot de gemeente het lerend vermogen van haar organisatie. Alle kennis die in pilots/projecten wordt opgedaan, kan zo gebundeld worden en kan de gemeente als organisatie de stap naar datagestuurd werken zetten. Deel na afronding van een pilot via de data academie opgedane kennis met de hele organisatie, communiceer openlijk.
- Regel bestuurlijk commitment bij elke pilot/project.
- Werk toe naar een centrale plek waaruit werken met zelflerende algoritmen en andere vormen van kunstmatige intelligentie aangestuurd en begeleid wordt.

# Pilot Geautomatiseerde data-analyse bijstandsuitkering

## Evaluatierapport

| | |
|---|---|
| Projectleider | Dino Zecic |
| Datum | 02-03-2021 |
| Opdrachtgever | Lambert Manden |
| Programmamanager | Hanneke Engels |
| Versie | 1.0 |
| Status | Definitief |

# Inhoudsopgave

# Aanleiding

Gemeenten voeren Participatiewet uit en betalen uitkeringen op grond van deze wet. Eén van de taken die gemeenten daarbij hebben, is ervoor zorgen dat uitkering rechtmatig wordt verstrekt. Dat het bij de inwoners terechtkomt die daar recht op hebben. Het rechtmatig verstrekken van uitkeringen legt de basis voor de solidariteit van ons systeem van sociale zekerheid.
Zowel door de lokale overheid als landelijk wordt hier altijd met veel belangstelling naar gekeken en over gediscussieerd.

Uit het regeerakkoord:
“Misbruik van sociale voorzieningen ondermijnt het draagvlak voor solidariteit. Het kabinet vindt het van belang dat uitvoerders, waaronder gemeenten, effectief gebruik maken van de mogelijkheden tot het delen, koppelen en analyseren van data, uiteraard met inachtneming van de geldende wettelijke regels en waarborgen. Dit kan ook uitkeringsgerechtigden helpen om regels na te leven en fouten te voorkomen.”

Ook gemeente Deventer heeft een actief handhavingsbeleid en heeft door de jaren verschillende vormen en methoden van handhaving toegepast. Van maandelijkse inkomstenverklaringen via principes van hoogwaardig handhaven en risicoscore berekeningen tot bestandsvergelijkingen en themacontroles, is gemeente Deventer altijd op zoek geweest naar een effectief en efficiënt handhavingsmodel.  Daarbij nooit te vergeten dat overgrote deel van inwoners met een uitkering geen slechte bedoelingen heeft en geen potentieel fraudeur is. Ook dat is een belangrijk uitgangspunt van een goed handhavingsmodel.
Met het bovenstaande in het achterhoofd en met inzicht dat hedendaagse technologie kansen biedt om het handhavingsmodel te verbeteren is gemeente Deventer in 2018 een pilot gestart met geautomatiseerde data-analyse door middel van een zelflerend algoritme.

Waarom een pilot?
Datagedreven werken, waarbij gebruik wordt gemaakt van geautomatiseerde data-analyse en een zelflerend algoritme, is nieuw voor onze organisatie. Om effecten van data-analyse-gestuurd werken te onderzoeken heeft de gemeente voor een pilot gekozen. Met de pilot willen we in de praktijk toetsen wat er voor nodig is om geautomatiseerde data-analyse op een verantwoorde manier in te zetten. Uitkomsten van een pilot zijn per definitie voor een deel onzeker, maar deze vernieuwende manier van werken kan alleen in de praktijk worden getoetst. Landelijk zijn er al ervaringen met data-analyse-gestuurd werken op het gebied van bijstand. Verschillende gemeenten experimenteren ermee of ontwikkelen deze manier van werken.

# Onderzoeksvraag

Centrale onderzoeksvraag luidt als volgt:
**Leidt geautomatiseerde data-analyse tot betere controle van bijstand?**

De centrale onderzoekvraag is in volgende deelvragen onderverdeeld:

- Zijn we door middel van data-analyse gestuurd werken in staat om effectiviteit van onze onderzoeken te verhogen?
- Zijn we door middel van data-analyse gestuurd werken in staat onze dienstverlening te verbeteren?
- Kan data-analyse gestuurd werken uitgebreid worden naar andere werkgebieden (bijvoorbeeld Wmo, schuldhulpverlening)?

# De pilot

De pilot is in 3 fases verdeeld:

1. Voorbereiding
2. Onderzoeken
3. Evaluatie

## Voorbereiding

De voorbereiding bestond uit volgende stappen:

1. Ophalen van extern juridisch advies gericht op randvoorwaarden waarbinnen de pilot uitgevoerd mag worden (bijlage 1)
2. Samenstellen van het projectteam die de selectie van de marktpartij en de uitvoering van de pilot zou gaan oppakken (bijlage 2)
3. Selecteren van de marktpartij via een aanbestedingsprocedure.
4. Samen met de geselecteerde partij is de pilot opgezet waar twee stappen in de voorbereidingsfase zijn opgenomen:
    a. Data-deep-dive. In deze stap is de kwaliteit van de gegevens beoordeeld. Aan het einde van deze stap was een GO/NO-GO moment ingebouwd.
    b. Modelleren van algoritme aan de hand van de door de gemeente aangeleverde gegevens. Om tot een algoritme te komen dat robuuste voorspellingen maakt, is de kwaliteit van de volgende 3 modeleringstechnieken getest: Random forest, Neurale netwerken en Lasso modeling. Op basis van deze test bleek een combinatie van de random forest en neurale netwerken het beste om de data te beschrijven. De combinatie van technieken zorgt er voor dat het algoritme op zowel de huidig bekende gegevens als de toekomstige gegevens van nieuwe bijstandsgerechtigden een robuuste voorspelling kan maken. De Lassomodellen voorspelden niet beter dan wanneer je random zou aanwijzen wie wel of niet fraude pleegt. Om die reden zijn de Lassomodellen in dit geval niet van toegevoegde waarde. De neurale netwerken en het Random forest daarentegen, voorspelden beter dan random. Om die reden is er voor een combinatie van deze twee modellen gekozen. Dit betekent dat 70% van de voorspellingen gebaseerd is op het Random forest en 30% van de voorspellingen gebaseerd is op Neurale netwerken. Door twee technieken te combineren, is het risico weggenomen dat één model alleen goed kan voorspellen tijdens het testen van de modellen, maar in de toekomst niet goed presteert. Meer informatie over de algoritme is in de technische documentatie van de algoritme terug te vinden (bijlage 3).

5. Sociale recherche heeft een onderzoeksprotocol opgesteld (bijlage 4)
6. Gebruik gemaakt van De Etische Data Assistent [1] van Universiteit Utrecht (DEDA).

## Onderzoek

Rekening houdend met randvoorwaarden die in het juridisch advies zijn meegegeven is de volgende fase van de pilot gestart.
De belangrijkste voorwaarden:

- verbod op koppelen van bestanden,
- verbod op geautomatiseerd nemen van besluiten,
- pseudonimiseren van gegevens en
- proportionaliteitsbeginsel van onderzoek.

Er zijn vier ronden van kansberekeningen geweest. Elke ronde nam circa 3 tot 5 maanden in beslag. Elke ronde bestond uit een aantal vaste stappen:

1. Aanleveren van gepseudonimiseerde gegevens aan Totta.
2. Kansberekeningen door algoritme.
3. Aanleveren van kansberekeningen aan gemeente.
4. Depseudonimiseren van gegevens door middel van de bij de gemeente bekende sleutel.
5. Onderzoek van de top 10 en de 10 laagst scorende dossiers. Gemeente Deventer heeft besloten om ook de 10 laagst scorende kansberekeningen te onderzoeken om zich op deze manier een beter beeld te vormen van de werking van de algoritme en om zogenaamde ‘profiling’ te voorkomen.
6. Onderzoek door Sociale Recherche, conform het opgestelde protocol. De kern van het protocol is de proportionaliteit van de aanpak. Eerst werd de kansberekening zelf op waarde beoordeeld, als een zgn. anoniem fraudesignaal. Pas wanneer deze beoordeling voldoende aanwijzingen opleverde werd er een uitgebreider onderzoek gestart, eerst administratief en dan, waar nodig, door middel van huisbezoek. Onderzochte bijstandsgerechtigde werd in kennis gesteld van het lopende onderzoek en het feit dat het onderzoek naar aanleiding van een kansberekening door een algoritme is gestart.
7. Invoeren van onderzoeksresultaten in het uitkeringssysteem.
8. Aanleveren van bijgewerkte gegevens aan Totta.
9. Een tussentijdse DEDA-werksessie om te bepalen of we nog steeds ethische vraagstukken goed hebben geborgd.

[1] De Etische Data Assistent van de Universiteit Utrecht: DEDA helpt data-analisten, projectmanagers en beleidsmakers om ethische problemen in dataprojecten, datamanagement en databeleid te herkennen.

# Resultaten

## Kwantitatief

| Ronde | Aantal hoogst scorende | In onderzoek genomen | Uitkering beëindigd | Score | Besparing | Bijzonderheden |
|---|---|---|---|---|---|---|
| **1** | 10 | 8 | 3 | 30% | € 42.600 | Naast 3 beëindigingen zijn twee dossiers door DWT opgepakt om traject richting arbeid weer te starten. |
| **2** | 10 | 3 | 0 | 0% | | |
| **3** | 10 | 2 | 1 | 10% | € 14.200 | En drie maanden uitkering teruggevorderd. |
| **4** | 10 | 8 | 0 | 0% | € 0 | |
| **Totaal** | 40 | 21 | 4 | 10% | € 56.800 | |

## Kwalitatief

Met deze pilot hebben we meer resultaten bereikt dan alleen het beëindigen van uitkeringen.
Overige resultaten:

- Vooruitlopend op conclusies is de algemene verwachting dat het werken met zelflerende algoritmen en andere vormen van kunstmatige intelligentie geen kwestie is van ‘of’ maar van ‘wanneer’
- Gemeente Deventer heeft met deze pilot kennis opgedaan op het gebied van werken met zelflerende algoritmen. Deze kennis kan gebruikt worden bij het verder ontwikkelen van datagedreven werken in de organisatie.
- Tijdens de pilot is gemeente Deventer in contact gekomen met andere gemeenten die ook hierover nadenken. Deze kennisbundeling is van toegevoegde waarde. Ten slotte is het vermelden waar dat gemeente Deventer in nauw contact is met mevrouw M. Wieringa [2]. Zij doet een promotieonderzoek naar de toepassing van algoritmen bij Nederlandse gemeenten. Haar onderzoek richt zich vooral op de transparantie en ethiek waarbij onze aanpak als een “goed voorbeeld” wordt gezien.
- Vanaf het begin van de pilot data-analyse was het evident dat een goede communicatie en openheid naar de burgers, die uit de analyse naar voren kwamen, belangrijk was.
  Om die reden is er dan ook voor gekozen om alle burgers die uit de analyse kwamen persoonlijk thuis te bezoeken en te informeren over het feit dat zij uit een data-analyse naar voren kwamen. Immers, een burger heeft het recht te weten waarom overheidsfunctionarissen zich aan de voordeur melden. In alle gevallen bleek dat de mededeling dienaangaande niet op weerstand stuitte. De meeste betrokkenen namen de mededeling voor kennisgeving aan en vonden het normaal dat zij gecontroleerd werden. Door geen van de betrokken burgers werd doorgevraagd op de details van de analyse. In 1 geval werd na deze mededeling zelf tevredenheid geuit omdat de betrokkene blij was dat de gemeente zich bij hem meldde. Deze persoon wilde namelijk graag aan het werk. Hij is actief

[2] Onderzoek van M. wieringa:
Titel: 'Approaching algorithmic account(-)ability: developing tools to foster formalized and practical transparency in municipal data projects'.
Promovenda: Maranke Wieringa, MA
Promotoren en supervisor: prof. dr. José van Dijck, prof. dr. Albert Meijer, dr. Mirko Tobias Schäfer
Loopduur: September 2018 t/m April 2023
Toolkit in ontwikkeling: Beraadslagingsinstrument voor Algoritmische Systemen (BIAS)

onder aandacht van het DWT gebracht. Resumerend kan worden gesteld dat het fenomeen data-analyse door de betrokken burgers positief is ontvangen.

## Financieel

Ook financieel heeft de pilot zich bewezen.

| | |
|---|---|
| **Opbrengsten** | € 56.800 |
| **Kosten** | € 41.500 |
| **Resultaat** | € 15.300 |

# Conclusie

Leidt geautomatiseerde data-analyse tot betere controle van bijstand?
Ja. Daarbij moet worden meegenomen dat deze manier van controle aanvullend moet zijn aan andere methoden en technieken w.o. fraudealertheid van medewerkers. Daarin verandert niets. Zich baseren op één controlemethode was en blijft gevaarlijk.

Daarnaast kan worden geconcludeerd dat:

- Er binnen de Participatiewet voldoende juridische kaders zijn waarbinnen het werken met algoritmen mogelijk is.
- Deze manier van controle van bijstand zich loont. De kosten zijn lager dan de baten. Het is natuurlijk niet altijd inzichtelijk te maken welke baten er zijn en welke financiële waarde ze vertegenwoordigen, maar inzetten van algoritmen kan in de basis tegen beperkte kosten.

Deelvragen

- Zijn we door middel van data-analyse gestuurd werken in staat om effectiviteit van onze onderzoeken te verhogen?
  Ja. Gemiddeld rendement is 10%. Dat is hoger dan bij bijvoorbeeld themacontroles, waar het percentage blijft steken op 3%. We onderzoeken n.a.v. een data signaal. Zonder deze data hadden we dit signaal niet. Dat is dus al een opbrengst. Vervolgens moeten we eigenlijk nog vrij traditioneel onderzoeken of er sprake is van fraude. Deze wijze van onderzoek verschilt nauwelijks/niet met die van regulier fraudesignalen. Voor de sociale recherche is er nauwelijks tijdswinst behaald, wel voor de kwaliteit van het bestand. Reguliere heronderzoeken laten een goed resultaat zien m.b.t. norm-mutaties en beëindigingen. Of de scorekans bij dataonderzoek hoger is kunnen we nu nog niet zeggen.
- Zijn we door middel van data-analyse gestuurd werken in staat onze dienstverlening te verbeteren?
  Ja. Door middel van zelflerende algoritmen kunnen we op een efficiënte manier grote hoeveelheden data analyseren en verbanden leggen die met ‘het blote oog’ niet te leggen zijn. Het hele principe is dat we onze dienstverleningsprocessen meer op maat kunnen maken. We controleren alleen daar waar het moet. Ook hier geldt dat blindstaren op dit soort hulpmiddelen potentieel gevaarlijk kan zijn. Onbedoeld kunnen profileren en uitsluiten onderdeel van onze processen worden, daar moeten we voor waken. Dit risico is goed te ondervangen door bijvoorbeeld steekproeven uit te voeren en door regelmatig wetenschappelijke hulpmiddelen toe te passen en in onze processen in te bouwen (zoals DEDA en nog te ontwikkelen Toolkit BIAS van Universiteit Utrecht). Er moet een systeem van ‘checks and balances’ zijn. Aan de andere kant is het controleren van bijstandsgerechtigden een plicht die altijd op gespannen voet staat met privacy en veiligheid van bijstandsgerechtigden. We moeten als overheid altijd beducht zijn op uitwassen die kunnen

voortkomen uit algoritmen of onderbuikgevoel van eigen medewerkers. Uitspraak van de rechtbank van 5 februari 2020 inzake het Systeem Risico Indicatie (SyRI) is een goede en noodzakelijke les daarin. Op deze manier toepassen van algoritmen is in strijd met artikel 8 van het Europees Verdrag voor de Rechten voor de Mens (EVRM).

- Kan de data-analyse gestuurd werken uitgebreid worden naar andere werkgebieden (bijvoorbeeld Wmo, schuldhulpverlening)?
  Ja. Deze conclusie is gebaseerd op inzicht dat we in deze casus over voldoende data van goede kwaliteit beschikken waarmee goede algoritmen gebouwd kunnen worden. Dan is het onderwerp van ondergeschikt belang. Uiteraard zal de kwantiteit en de kwaliteit van data per casus van tevoren moeten worden bepaald. Nu hebben we gewerkt met kansberekeningen achteraf op lopende dossiers. Het is voor te stellen dat kansberekeningen ook vooraf aan het begin van het proces ingezet kunnen worden. Bijvoorbeeld hoe groot is de kans dat een schuldhulpverleningstraject slaagt, gelet op de kenmerken van die aanvraag. Mocht de kans klein zijn, dan zijn we in staat om op tijd in te grijpen zodat de kans groter wordt. Maar kunnen zien waar de fraude zich af zou kunnen spelen en dat duiden zou ook meerwaarde hebben.

# Advies

Wij hebben een nuttig pilot uitgevoerd, met mooie resultaten en belangrijke conclusies. Maar zoals uit voorgaande blijkt, is er nog veel te leren. Juridisch en procesmatig zijn algoritmen, met juiste waarborgen, goed toepasbaar. Waar nog onduidelijkheid over is, is de maatschappelijke en bestuurlijke verantwoording die moet kunnen gegeven worden. Vragen die daarbij spelen:

- Wat moet ik als gemeente geregeld hebben om zelflerende algoritmen onderdeel van de reguliere werkprocessen te maken?
- Kan ik als gemeente voldoende uitleggen hoe een algoritme werkt?
- Kan ik als gemeente signaleren wanneer er een 'bias' optreedt en hoe voorkom ik het juist?
- Hoe ga ik als gemeente om met het negatieve imago die algoritmen nu hebben, terwijl ik weet dat ze toegevoegde maatschappelijke waarde kunnen hebben?

De gemeente staat aan het begin van deze ontwikkeling en gelukkig blijkt er voldoende interesse zowel lokaal als landelijk en vanuit de wetenschap om algoritmen op een verantwoorde manier te inzetten. Bovendien dient de taboe op inzet van algoritmen in een maatschappelijk/politiek debat doorbroken te worden.

Met het bovenstaande in het achterhoofd adviseren we gemeente Deventer het volgende:

1. Maak werken met zelflerende algoritmen vast onderdeel van het signaal gestuurd werken binnen de Participatiewet. Juridisch kader maakt het mogelijk, er is voldoende data van goede kwaliteit en er is voldoende kennis bij de medewerkers om binnen de kaders van de wet en ethiek met geautomatiseerde data-analyse te werken.
2. Onderzoek verder door pilots/projecten uit te voeren met verschillende onderwerpen en wettelijke kaders.
   Bij het organiseren van een pilot neem volgende kritische succesfactoren mee:
   a. Juridisch kader is van tevoren vastgesteld.
   b. Ethische overwegingen en kaders zijn van tevoren bekend (DEDA).
   c. Formuleer een onderzoekvraag. Onderzoeksvraag richt zich niet alleen op het inhoudelijke onderwerp, maar ook op andere randvoorwaardelijke zaken.
      Zoals: is onze data van goede kwaliteit, hebben we de juiste competenties om met zalflerende algoritmen te werken, etc.
   d. Zorg voor een multidisciplinaire projectgroep, waar inhoudelijk betrokken collega's worden aangevuld met collega's van ICT, Kennis en Verkenning, Privacy.
   e. Betrek project Datagedreven werken bij de volgende pilots/projecten. Op deze manier vergroot de gemeente het lerend vermogen van haar organisatie. Alle kennis die in pilots/projecten wordt opgedaan, kan zo gebundeld worden en kan de gemeente als organisatie stap naar data gestuurd werken zetten. Deel na afronding van een pilot via de data academie opgedane kennis met de hele organisatie, communiceer openlijk.
   f. Regel bestuurlijk commitment bij elke pilot/project.
   g. Werk toe naar een centrale plek waaruit werken met zelflerende algoritmen en andere vormen van kunstmatige intelligentie aangestuurd en begeleid wordt.
3. Binnen Deventer zijn er ontwikkelingen waarin geautomatiseerde data-analyse goed past, als mogelijk vervolg. Kenmerken van deze projecten zijn (afgeleid van en aansluitend op de succesfactoren):
   a. Helder juridisch kader wat betreft rechtmatig, proportionaliteit en doelbinding.
   b. Voldoende data beschikbaar om te kunnen leren (of data die beschikbaar komt)
   c. Scherp beeld bij de mogelijke opbrengsten in uitvoering van het (wettelijke) proces, financieel en/of bediening van de inwoners.

   Dit komt neer op de volgende initiatieven waarvan in meer of mindere mate al gedacht wordt aan datagestuurd werken:

    d. Voorspelling van de Wmo budgetten; analyseren van factoren die de budgetten beïnvloeden.
    e. DeltaView ontwikkeling waarin informatie samengekomen is waarmee schulden preventie mogelijk wordt
    f. Onderzoeken of algoritme kan werken bij fraudedetectie bij Wmo
4. Trek samen op met andere gemeenten, wetenschap en landelijke organisaties.

# Bijlagen

1. Juridisch advies Eiffel Legal Office

2. Technische documentatie algoritme

3. Onderzoeksprotocol

# ADVIES

# FRAUDEONDERZOEK

## GEMEENTE DEVENTER

EIFFEL
LEGAL
FINANCE
PROCESS

| | |
|---|---|
| *Opdrachtgever* | ***Gemeente Deventer*** |
| *Paraaf voor akkoord* | |

| | |
|---|---|
| *Auteur* | ***Mr. Stefanie Kelterman*** |
| *Datum* | ***dinsdag 11 juli 2017*** |
| *Versie* | ***1.0*** |
| *Bijlage(n)* | ***Neen*** |

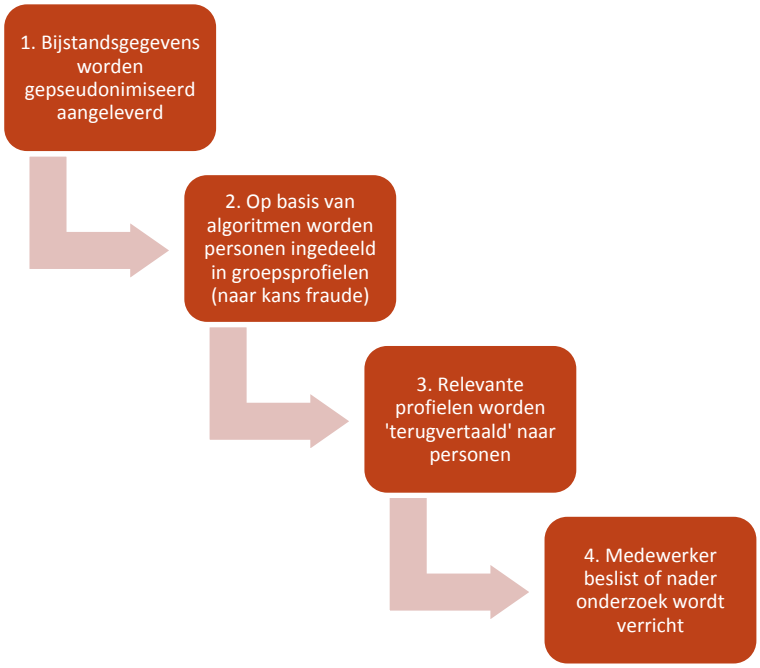
## ADVIESVRAAG

De gemeente Deventer heeft het Legal Office van EIFFEL gevraagd advies uit te brengen over een voorgenomen pilot om de rechtmatigheid van verleende bijstand te controleren en zo fraude beter op te sporen. Dit advies beschrijft de mogelijkheden die de privacywetgeving – waaronder ook begrepen de Algemene Verordening Gegevensbescherming (AVG) – biedt om deze pilot uit te voeren.

## BESCHRIJVING PILOT

De pilot die de gemeente Deventer voornemens is uit te voeren heeft tot doel om de algemene steekproefsgewijze controle op mogelijke bijstandsfraude te vervangen door een controle binnen bepaalde risicoprofielen. De verwachting is dat door controles uit te voeren binnen de vooraf vastgestelde risicoprofielen het slagingspercentage van de fraudeopsporing wordt vergroot. De fraudeprofielen worden vastgesteld op basis van ervaringen met fraudebestrijding binnen de gemeente.

De toekenning van personen aan deze profielen wordt gedaan in samenwerking met een externe partij. Welke partij dit wordt is ten tijde van het schrijven van dit advies nog niet bekend.




1. De gemeente levert relevante gegevens aan van bijstandsgerechtigden, waarbij de identificerende gegevens als naam en bsn zijn vervangen door een klantnummer.
2. Deze gegevens worden geanalyseerd en hieraan worden kenmerken met wegingen toegevoegd. Door middel van algoritmes wordt het bestand verdeeld in groepsprofielen. Deze groepsprofielen zien op een percentuele mogelijkheid dat bijstandsfraude wordt gepleegd.
3. Door de derde partij worden de resultaten terug geleverd aan de gemeente. De gemeente vertaalt deze resultaten terug naar personen;
4. Op de profielen met de hoogste percentages wordt nader onderzoek verricht, waarbij eerst gekeken wordt in het dossier of er ook andere aanwijzingen zijn om onderzoek te doen. De beslissing om onderzoek te doen ligt altijd bij een medewerker van de gemeente.

## TOEGEPAST JURIDISCH KADER

### Grondslag voor verwerking

De huidige én toekomstige privacywetgeving schrijft voor dat verwerking van persoonsgegevens niet plaatsvindt, tenzij hiervoor een grondslag bestaat. Deze grondslagen staan limitatief beschreven in de wet. Onderzoek doen naar de juiste toekenning van bijstandsgelden is een publieke taak die de gemeente opgelegd heeft gekregen in de Wet Werk en Bijstand. In artikel 18 van die wet staat dat de gemeente de bijstand moet afstemmen:

*"Het college stemt de bijstand en de daaraan verbonden verplichtingen af op de omstandigheden, mogelijkheden en middelen van de belanghebbende."*

In artikel 41, 10e lid wordt de gemeente de mogelijkheid geboden om ten behoeve van die afstemming onderzoek te doen:

"*In de aanvraag verleent belanghebbende het college een machtiging om onderzoek in te stellen naar de juistheid en volledigheid van de verstrekte gegevens en zo nodig naar andere gegevens die noodzakelijk zijn voor de verlening dan wel de voortzetting van bijstand*."

Het gebruik van persoonsgegevens voor het doen van onderzoek naar bijstandsfraude is daarmee toegestaan. Dit onderzoek op de hierboven beschreven wijze, waarbij gebruik wordt gemaakt van algoritmes en profielen, is aan enkele aanvullende voorwaarden gebonden. Deze worden hier besproken.

### a) Doelbinding

De gegevens die de gemeente gebruikt om de analyse te laten uitvoeren zijn beperkt tot gegevens:

- die zijn verzameld voor het doel 'verlening bijstand'. Er mogen dus geen andere gegevens worden betrokken die voor een ander doel zijn verkregen door de gemeente of een andere instantie. Er mag dus geen verrijking plaatsvinden van de gegevens. Gegevens van het Inlichtingenbureau mogen niet gebruikt worden als de gemeente hierover zelf al niet de beschikking had in zijn bijstandstaak;
- waarvan vooraf is bepaald dat die relevant zijn voor het bepalen van het frauderisico. Er mogen dus niet willekeurig gegevens in de analyse worden gebruikt om te zien of die eventueel resultaat opleveren. Wel is het mogelijk als het bij de gemeente bekend wordt dat een persoon gescheiden is, dit in de volgende analyse mee te nemen (als voorafgaand aan de analyse in algemene zin is bepaald dat het gegeven 'pas gescheiden' relevant is voor fraudeonderzoek).

### b) Proportionaliteit

Het gebruik van persoonsgegevens voor het doen van onderzoek naar bijstandsfraude vormt een inbreuk op de privacy van de betrokkenen. Deze inbreuk wordt gerechtvaardigd door het nagestreefde doel, zolang geen sprake is van een onevenredige belasting. Van belang is daarbij de frequentie van analyse. Als viermaal per jaar een dergelijke analyse wordt gemaakt is geen sprake van een onevenredige belasting. Als echter de gegevens doorlopend geanalyseerd worden en telkens als een wijziging in de gegevens plaatsvindt die relevant kan zijn voor de profilering van die persoon een signaal naar de gemeente gaat, kan wel sprake zijn van onevenredige belasting.

### c) Automatische beslissingen

De inzet van analyses op basis van algoritmes en profiling mag nooit leiden tot geautomatiseerde besluitvorming. Leg dus duidelijk vast in het proces dat altijd door een medewerker van de gemeente wordt beslist:

1. of nader onderzoek wordt verricht naar een persoon die in een bepaald profiel valt, en
2. of de resultaten van dat onderzoek leiden tot aanpassing van de bijstand.

## AANDACHTSPUNTEN

- **Privacy Impact Assessment**: onder de AVG wordt het verplicht om in het geval van een verwerking met een verhoogd risico (wat profilering over het algemeen is) een Privacy Impact Assessment (PIA) te doen. Vooruitlopend op die verplichting adviseer ik om dat voor deze pilot ook te doen. Met het uitvoeren van een PIA leg je aantoonbaar vast:
  - dat je voorafgaand aan de verwerking hebt nagedacht over privacy risico's (privacy by design) en hierop ook maatregelen hebt getroffen;
  - welke afwegingen je hebt gemaakt in het kader van de met de bewerker te delen persoonsgegevens (welke gegevens zijn noodzakelijk om een goede analyse te maken);
  - dat geen sprake is van geautomatiseerde beslissingen op basis van de profielen;
  - dat de verantwoordelijkheid voor het fraudeonderzoek volledig binnen de gemeente ligt.

  Een goed ingevulde en gedocumenteerde PIA voorkomt vragen van de toezichthouder of een auditeur en zorgt voor borging van bepaalde beleidskeuzes.
- **Verplichting tot teruglevering resultaten**: het verbeteren van de indeling in profielen gebeurt door de resultaten van de onderzoeken (werd daadwerkelijk gefraudeerd?) toe te voegen aan de eerdere analyse. Let hierbij goed op welke verplichting je hierover aangaat met de derde partij. Het ligt voor de hand dat de derde partij die resultaten niet alleen wil gebruiken voor het verbeteren van de algoritmes voor de gemeente Deventer, maar voor al zijn klanten. Hiervoor mogen geen persoonsgegevens van de gemeente worden gebruikt. Let hierbij dus op de herleidbaarheid van de resultaten. In principe kunnen wel percentages worden teruggeleverd (18% uit het profiel 'hoog frauderisico' heeft daadwerkelijk fraude gepleegd), tenzij de aantallen zo klein zijn dat er risico is op spontane herkenning (50% van een profiel van 2 personen is redelijk eenvoudig te herleiden);
- **Bepalen profielen**: een analyse van bijstandsgegevens biedt eindeloos veel profielmogelijkheden. Naast de praktische kant (een te grote verscheidenheid aan profielen leidt niet tot vereenvoudiging van de opsporing) is hier ook een privacykant om rekening mee te houden. De vooraf bepaalde profielen moeten direct verband houden met de opsporingstaak. Een profiel moet dus onderscheidend zijn in het risico op fraude (denk aan hoog, middel, laag risico). Het profiel mag niet onderscheidend zijn op onderwerpen die niet direct te maken hebben met fraude-opsporing (denk aan 'hoog risico op fraude én heeft aanvraag WMO lopen').
- **Doel van de analyse**: partijen die analyses aanleveren bieden vaak nog veel meer mogelijkheden van die analyses aan om het product aantrekkelijker te maken. Dit advies ziet uitsluitend op het gebruik van de analyse ter ondersteuning van het fraudeonderzoek in de bijstand. Voor iedere nieuwe toepassing van de analyse moet opnieuw bekeken worden of deze de publieke taak ondersteunt van de gemeente en hiervoor ook noodzakelijk is. In een van de toegezonden presentaties kwam ik tegen dat de analyse ook erg handig is om te gebruiken ter voorbereiding van een zogenaamd 'keukentafelgesprek'. Zonder een diepgaande juridische analyse te doen lijkt mij dat de inbreuk op de privacy door voorafgaand aan een verkennend gesprek een complete analyse te doen niet gerechtvaardigd wordt door het doel.

## BEWERKERSOVEREENKOMST

Als de analyse van de bijstandsgegevens wordt uitbesteed aan een derde partij, moet met die partij een bewerkersovereenkomst worden afgesloten die ziet op de uitwisseling van en de omgang met persoonsgegevens.

Ook als identificerende gegevens al bij de gemeente zijn vervangen door een klantnummer, is de kans op herleidbaarheid groot. De groep bijstandsgerechtigden binnen één gemeente is immers niet groot. Daarbij kan het gegevensbestand dat van één persoon wordt aangeleverd zoveel gegevens bevatten dat spontane herkenning mogelijk is. Het is daarom belangrijk om goede afspraken te maken over het gebruik van de gegevens.

Naast de afspraken over de wijze van gegevensuitwisseling en het doel van de uitwisseling is het belangrijk om in ieder geval de volgende onderwerpen te benoemen:

- Datalekken: als er gegevens lekken is de gemeente als verantwoordelijke verplicht om dit lek binnen 72 uur te melden bij de Autoriteit persoonsgegevens. Ook als het lek plaatsvindt bij de bewerker. De bewerker moet daarom verplicht worden om direct bij ontdekking van een datalek de gemeente daarvan op de hoogte te brengen;
- Geheimhouding: nu de kans op spontane herkenning aanwezig is en bijstandsgegevens gevoelige gegevens zijn, is het van belang dat de medewerkers van de bewerker een geheimhoudingsverklaring tekenen en dat het aantal mensen dat de gegevens inziet beperkt blijft;
- Vernietiging van gegevens: de gegevens die de gemeente aanlevert worden alleen gebruikt voor de eenmalige analyse, direct daarna worden de gegevens vernietigd. Let hierbij op wat de derde partij voorstelt met betrekking tot het zogenaamd ‘zelflerend vermogen’ van de algoritmes. Worden gegevens hiervoor bewaard? Uitgangspunt van de privacywetgeving is dat gegevens niet langer worden bewaard dan nodig voor de taak. Het bewaren van gegevens om ooit te gebruiken voor het ‘slimmer’ maken van de algoritmes is daarom niet toegestaan.

## ONDERZOEK DOOR GEMEENTE

Tot slot wijs ik op een uitspraak van de Centrale Raad van Beroep (ECLI:NL:CRVB:2014:2947) waarin de Raad heeft bepaald dat de gemeente haar kerntaken bij de uitvoering van de bijstand, zoals de preventie van bijstandsfraude, niet mag uitbesteden aan een commercieel bedrijf. Deze kerntaken dienen binnen het publieke domein te worden uitgevoerd.

In deze uitspraak was het fraudeonderzoek zelf binnen het sociaal domein uitgevoerd door een commerciële partij. Deze partij kreeg betaald op basis van ‘no cure no pay’ en de medewerkers van deze partij hadden zich ten onrechte uitgegeven voor handhavingsmedewerkers van de betreffende gemeente. De Raad komt tot de conclusie dat het bewijs dat op die manier verworven was niet gold als rechtmatig verkregen bewijs en daarom niet kon worden meegenomen.

De pilot zoals voorgenomen door de gemeente Deventer heeft een andere insteek, met name doordat het daadwerkelijke fraude-onderzoek wordt uitgevoerd door medewerkers van de gemeente. Het toont echter wel aan dat het belangrijk is om dat vast te leggen en ook duidelijk vast te leggen dat de in te schakelen derde partij geen baat heeft bij meer of minder potentiële fraudeurs.

# PROTOCOL SOCIALE RECHERCHE PILOT DATA-ANALYSE I.O.

**Werkwijze afhandeling potentiële fraudesignalen uit data-analyse**:

De casussen die aangeleverd worden uit de data-analyse zullen volgens de reguliere werkwijze van de sociale recherche Deventer afgehandeld worden.
Dit impliceert dat daarbij alle wettelijke kaders gehanteerd zullen worden, waarbij nadrukkelijk bij elke aangeleverde casus gekeken wordt naar proportionaliteit en subsidiariteit.

**Gehanteerde wettelijke kaders:**

| | |
|---|---|
| Artikel 8 van het EVRM | (schending privacy) |
| AVG | (privacy waarborgen) |
| Artikel 17 van de Participatiewet | (Inlichtingenplicht) |
| Artikel 53a van de Participatiewet | (onderzoeksbevoegdheid) |
| Artikel 1 van Protocol 12 bij het EVRM | (verbod discriminatie) |

Per casus zal door een sociaal rechercheur beoordeeld worden of er sprake is van een redelijke grond om een nader onderzoek in te stellen.

Een onderzoeksmiddel dat inbreuk maakt op het privéleven mag slechts worden ingezet, indien dit noodzakelijk is in het belang van een van de in artikel 8 EVRM genoemde doelen en dus voldaan is aan de vereisten van proportionaliteit en subsidiariteit die de inbreuk rechtvaardigen. De inbreuk die de onderzoekers op het privéleven van bijstandsgerechtigden maken, door per casus te beoordelen of er sprake is van uitkeringsfraude, en eventueel een nader onderzoek in te stellen, dient niet onevenredig zwaar te zijn in verhouding tot het te dienen doel, namelijk de bestrijding van misbruik van socialezekerheidsregelingen.

Naar de betreffende belanghebbenden zal transparant worden gecommuniceerd over de data-analyse. Als deze aanleiding is geweest tot het onderzoek zal dit naar de belanghebbende als een signaal voorgekomen uit data-analyse worden benoemd.

**Proces:**
Concreet zal de volgende werkwijze worden gevolgd:

1. Binnenkomst signaal data-analyse.

2. Selectie van 10 hoogst scorende dossiers.
   Op deze wijze controleren we de precisie van het algoritme.

3. Selectie van 10 laagst scorende dossiers.
   Op deze wijze controleren we de precisie van het algoritme en brengen we evenwicht in het onderzoek. Voeren we de controle uit alleen op de hoogst scorende dossiers dan bestaat de kans dat de “profiling” het doel wordt en niet de controle van de rechtmatigheid. Kansberekeningen door middel van een algoritme zijn ondersteunend en niet leidend.

4. Wanneer blijkt dat een dossier uit de top-10 al in onderzoek is, dan nemen we het eerstvolgende dossier uit top-20 in onderzoek.

5. Beoordeling kwaliteit signaal door sociaal rechercheur door controle bronmateriaal.

6. Als kwaliteit is gecontroleerd, bepalen welke onderzoeksmethode dient te worden toegepast.

7. Onderzoeksmethode (niet limitatief: dossieronderzoek, opvragen nadere informatie bij klant, opvragen informatie bij andere instanties, steekproefsgewijze heimelijke waarnemingen, heimelijke waarnemingen, huisbezoek) toepassen.

8. Klant informeren (notificatieplicht).

9. Terugkoppeling naar Totta over kwaliteit signaal.

10. Bij alle geselecteerde dossiers worden werkprocessen in het systeem gestart en binnen drie maanden met een resultaat afgesloten. Deze gegevens zijn nodig voor het verrijken van data. Met de verrijkte data wordt de volgende set kansberekeningen opgeleverd.
    Hierin zit het lerende aspect voor zowel de organisatie als voor het algoritme.

11. Afhankelijk van de resultaten uit de eerste set van kansberekeningen wordt bij de volgende set besloten welke dossiers onderzocht worden en op welke wijze.

12. Het bovenstaande wordt maximaal 3 keer gedaan.

Resumerend wordt een casus uit de data-analyse op eenzelfde wijze als een binnengekomen signaal uit een andere bron behandeld.

# Technische documentatie Voorspelmodel

## Onrechtmatigheidscontrole

gemeente Deventer

**TOTTA DATA LAB**

CONNECTING THE DOTS

# 1 Introductie

Totta data lab heeft een algoritme ontwikkeld voor de gemeente Deventer. Dit algoritme brengt patronen in data in kaart die duiden op een verhoogd risico op onrechtmatigheden. Op basis van deze patronen voorspelt het algoritme voor elke cliënt de kans op onrechtmatig gedrag. In dit document is uitgelegd welke data gebruikt wordt en hoe dit algoritme tot stand is gekomen.

## 1.1 Data

Alle databronnen die verwerkt zijn in het model, zijn in overleg en na goedkeuring van de gemeente Deventer gebruikt. Van huidige en voormalige bijstandsgerechtigden in de gemeente Deventer zijn de uitkeringsgegevens gegevens omtrent samengevoegd met:

- achtergrondkenmerken van cliënten;
- kinderen;
- vakantieperioden;
- blokkades;
- rechten, plichten en ontheffingen;
- participatietrajecten;
- bijzondere situaties;
- vermogen;
- werkprocessen;
- controles en fraude registratie;
- debiteuren.

Aan deze set gegevens is een label toegevoegd dat aangeeft of een bijstandsgerechtigde in het verleden onrechtmatigheden heeft gepleegd. Dit label is toegekend wanneer bijvoorbeeld na een controle op fraude inderdaad onrechtmatigheden zijn gevonden, maar ook wanneer er vorderingen zijn gedaan vanwege fraude in de wet werk en bijstand of in de bijzondere bijstand.

Uiteindelijk willen we een robuust model ontwikkelen, dat zo correct mogelijk fraude voorspelt. Met robuust bedoelen we dat de voorspellingen van het model weer net zo goed zijn wanneer het model voorspellingen doet op een nieuwe dataset. Wanneer het model in onze test 50% fraude correct voorspelt, maar bij een nieuwe dataset maar 20% correct voorspelt, dan is het geen robuust model. Voor het vinden van een robuust algoritme is deze data set willekeurig verdeeld in een training dataset (70% van de data) en een test dataset (30%). Daarnaast zijn alleen personen geselecteerd die na 2013-01-01 een uitkering hebben. Vanaf dat moment is voor het eerst onrechtmatig gedrag geregistreerd. Door in dezelfde periode gedrag cliënten mee te nemen, zijn de uitkeringen beter met elkaar te vergelijken. De training dataset is gebruikt om aan de hand van verschillende modeleringstechnieken patronen te herkennen in de data. De test dataset is gebruikt om te valideren of de gevonden patronen robuust genoeg zijn om de data te kunnen beschrijven. Bij het valideren van de verschillende modeleringstechnieken is de kwaliteit van een voorspelling vergeleken met een model dat een willekeurige voorspelling maakt. Daarnaast is er gekeken naar de 10 gevallen met de hoogste voorspelde kans op onrechtmatigheid en vervolgens is vergeleken hoeveel gevallen daarvan daadwerkelijk onrechtmatig gedrag vertonen. Een combinatie van de geteste modelleertechnieken resulteert in een eindmodel. De technieken dragen niet allemaal evenveel bij aan het eindmodel, maar de bijdrage van de technieken is afhankelijk van de prestaties gedurende de tests. Des te meer fraude gevallen in de top 10 correct zijn voorspelt in de tests, des te meer de techniek zal meewegen in het eindmodel. Met het

uiteindelijke model zijn vervolgens voorspellingen gemaakt op de voorspel dataset. In de voorspel dataset zitten alleen personen die op dat huidige moment in de uitkering zaten.

## 1.2 Algoritme

Om tot een algoritme te komen dat robuuste voorspellingen maakt, is de kwaliteit van de volgende modeleringstechnieken getest: Random forest, Neurale netwerken en Lasso modeling. Van deze 3 technieken bleek een combinatie van de Random forest en Neurale netwerken het beste om de data te beschrijven. Er is voor een combinatie van twee modellen gekozen: Random forest en neurale netwerken. Dit is om ervoor te zorgen dat het algoritme op zowel de huidig bekende gegevens als de toekomstige gegevens van nieuwe bijstandsgerechtigden een robuuste voorspelling kan maken. De Lassomodellen voorspelden niet beter dan wanner je random zou aanwijzen wie wel of niet fraude pleegt. Om die reden zijn de Lassomodellen in dit geval niet van toegevoegde waarde. De neurale netwerken en het Random forest daarentegen, voorspelde beter dan random. Om die reden is er voor een combinatie van deze twee modellen gekozen. Dit betekent dat 70% van de voorspellingen gebaseerd is op het Random forest en 30% van de voorspellingen gebaseerd is op Neurale netwerken. Door twee technieken te combineren, is het risico weggenomen dat één model alleen goed kan voorspellen tijdens het testen van de modellen, maar in de toekomst niet goed presteert.

# 2 Uitleg van het project

Totta data lab heeft voor de gemeente Deventer een algoritme ontwikkeld dat de kans op fraude in de bijstand voorspelt voor elke bijstandsgerechtigde. Dit algoritme brengt patronen in data in kaart die duiden op een verhoogd risico op fraude. Het algoritme genereert een kansberekening op onrechtmatigheden op cliëntniveau. Op basis van deze kansberekeningen voert de gemeente onrechtmatigheidsonderzoeken uit. Tot op heden blijkt deze aanpak succesvol, aangezien er in 2018 en 2019 meerdere onrechtmatigheden zijn gevonden.

In dit document wordt uitgelegd hoe het algoritme werkt. Eerst zal er gekeken worden naar de data die gebruikt is voor het ontwikkelen van een voorspellend model. Vervolgens zal worden uitgelegd hoe deze data verwerkt en geprepareerd is. Tenslotte laten we zien hoe is onderzocht welke techniek de data het beste kan beschrijven zodat uiteindelijk een robuuste voorspelling gemaakt kan worden.

## 2.1 Definitie van een algoritme

Voordat uitgelegd gaat worden hoe het algoritme van Deventer werkt, bespreken we eerst de definitie van een algoritme. Een algoritme is een eindige reeks handelingen die vanuit een begintoestand naar een beoogd doel leidt. Een algoritme is vergelijkbaar met een recept waarbij eieren, meel en melk de begintoestand zijn. Na een reeks van handelingen worden deze ingrediënten omgevormd tot een pannenkoek. Voor het algoritme van Deventer zijn de gegevens van bijstandsgerechtigden gebruikt als begintoestand. Deze gegevens zijn na verschillende handelingen, gebruik makende van modeleertechnieken, omgevormd tot een kans op onrechtmatigheden bij bijstandsgerechtigden.

# 3 De data

Voor het voorspellen van onrechtmatigheden bij bijstandsgerechtigden binnen de gemeente Deventer zijn verschillende databronnen gebruikt. In dit hoofdstuk wordt uitgelegd welke gegevens nodig zijn om een voorspelling te maken. Alle databronnen verwerkt in het model, zijn in overleg en na goedkeuring van de gemeente Deventer gebruikt

## 3.1 Brondata

De data die wij nodig hebben bestaat uit gegevens omtrent:

1. Uitkeringsdossiers
2. Cliëntgegevens
3. Kinderen
4. Vakanties
5. Blokkades
6. Rechten, plichten en ontheffingen
7. Participatietrajecten
8. Bijzondere situaties
9. Vermogen
10. Werkprocessen
11. Controles en fraude registratie
12. Debiteuren

Niet alle aangeleverde gegevens zijn uiteindelijk meegenomen in het model. Dit komt omdat de we alleen een goed werkend model kunnen ontwikkelen op data met voldoende kwaliteit van voldoende kwantiteit.

## 3.2 Bevoordeeld model

Een bevooroordeeld model ontstaat wanneer het model gestuurd wordt door alleen vroegere controles. Hierdoor zouden personen die al eerder gecontroleerd zijn, weer opnieuw gecontroleerd worden. Om dit te voorkomen, nemen we de resultaten van de nieuwe fraude controles mee in het model. Hierdoor leert het model nog beter welke patronen wel een indicatie van fraude zijn en welke niet.  Naast het meenemen van de nieuwe controle resultaten elk kwartaal, voorkomen we de bias ook met combinaties van modellen. Verschillende modelleertechnieken zoals Random forest of neurale netwerken detecteren fraude patronen op een andere manier. In deze technieken wordt fraude op een andere manier getriggerd. Door de voorspellingen van meerdere technieken te combineren, zorgen we verschillende fraude patronen in kaart hebben. Hiermee voorkomen we ook deels dat dezelfde personen opnieuw naar boven komen met de hoogste kans op fraude. Als laatste, ontwikkelen we alleen modellen op basis van data die van voldoende kwaliteit en kwantiteit is. Wat dit precies inhoudt lees je in het volgende hoofdstuk.

## 3.2 Data kwaliteit & kwantiteit

Wanneer de data niet van voldoende kwaliteit en kwantiteit is, zal je model niet goed presteren. Je zult dan vergelijkbare voorspellingen krijgen als wanneer je random zou aanwijzen wie fraude pleegt.

### 3.3.1 Vullingsgraad

De data moet voldoende gevuld zijn. Wanneer variabelen voor 80% missende waarden hebben, dan is dit niet werkbaar. Dit schetst geen goed beeld van de werkelijke situatie want je weet maar van 20% van de personen wat de situatie is. We hanteren voor dit

algoritme een minimale vulling van 50%. Dit baseren we op de ervaringen die we hebben op gedaan bij het ontwikkelen van eerdere voorspelmodellen.

### 3.3.2 Variatie

De data moet voldoende variatie laten zien. Wanneer een variabele alleen gevuld is met dezelfde categorie, dan is er geen variatie. Stel dat iedereen in je dataset een man is, dan kan het algoritme geen onderscheid maken op basis van geslacht. Omdat geslacht nooit verschilt, is dit geen bruikbare variabele voor het algoritme. Daarnaast mag een variabele ook niet te veel variatie hebben. Stel dat de variabele 'reden van bijstandsbeëindiging' voor iedereen anders zou zijn. Dan zou het algoritme veel te weinig voorbeelden hebben van elke situatie om een duidelijk patroon te herkennen. Er moeten dus genoeg voorbeelden zijn van elke categorie om van te leren. Om die reden gebruiken wij alleen categorische variabelen waarbij er in elke categorie minimaal ongeveer 5% van de bijstandscliënten voorkomt.

### 3.3.3 Betrouwbaarheid

De data moet voldoende betrouwbaar zijn om mee te kunnen werken. Zo is het belangrijk dat wijzigingen in gegevens consequent worden geüpdatet. Wanneer de gegevens niet consequent zijn geüpdatet, kun je er niet van op aan dat de status van een cliënt op dat moment klopt. Wanneer de status niet klopt, kun je geen goed beeld schetsen van de werkelijke situatie. Dit is wel nodig om een betrouwbare voorspelling te kunnen maken. Naast het consequent updaten van de gegevens is het ook belangrijk dat iedereen op dezelfde manier categorieën van variabelen vult. Wanneer niet iedereen het gedrag van een klant hetzelfde beoordeelt, wordt er ook geen betrouwbaar beeld geschetst.

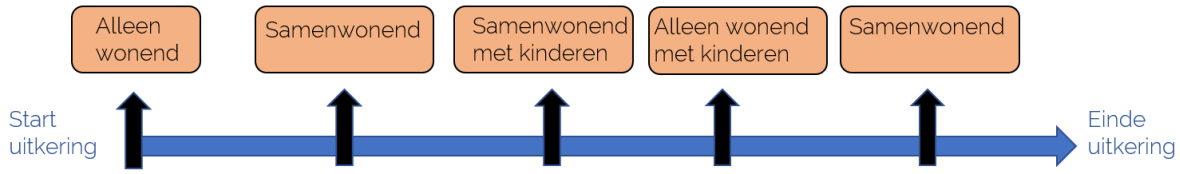
### 3.3.4 Paaseieren

Het kan gebeuren dat er in variabelen indicaties worden gegeven of iemand gecontroleerd is op fraude of dat iemand een onrechtmatigheid heeft begaan. In principe staan deze gegevens in de fraude controle tabel of in een debiteuren tabel, maar soms kunnen ze ook in andere tabellen voorkomen.. Dit soort variabelen nemen we nooit mee in het model. Als je dit namelijk wel doet, dan vertel je eigenlijk van tevoren aan het model dat iemand onrechtmatig gedrag heeft gepleegd. Dat is dan geen voorspelling meer, want je weet al van tevoren dat er sprake is van onrechtmatigheid. Soms is het niet meteen duidelijk dat een variabele onrechtmatigheid indiceert. We nemen zo'n variabele dan wel mee in het model, maar komen er dan tijdens het testen van het model achter dat het model alle fraudegevallen correct voorspelt. Als dat gebeurd, dan weten we dat het mode van tevoren al wist wie er fraude heeft gepleegd en dan weten we dat er een variabele is die deze informatie aan het model geeft. Deze variabele noemen we dan een paasei. We moeten dan namelijk

### 3.3.5 Historische data

Wanneer er geen veranderingen van gedrag zijn geregistreerd over de tijdlijn (Figuur 1) van de uitkering, is het moeilijk om individuele gedragspatronen te herkennen waarop onrechtmatigheden voorspelt kunnen worden. Stel dat iedereen hetzelfde gedrag zou vertonen aan het begin van de uitkering en aan het einde van de uitkering, dan zouden er niet zoveel gegevens zijn om op te bepalen waarom de een wel onrechtmatigheden pleegt en de andere niet. Wanneer gedrag verandert dan is het mogelijk om een bepaald omslagpunt gedurende de uitkering aan te wijzen dat een indicatie voor onrechtmatigheid zou kunnen zijn.

Figuur 1 - Voorbeeld van historische data:




### 3.3.6 Databronnen en koppelsleutels

Om met de data te kunnen werken moeten er koppelsleutels aanwezig zijn waarop de verschillende databronnen aan elkaar gekoppeld kunnen worden. Ook voor de koppelsleutels geldt dat ze voldoende gevuld moeten zijn. Soms kan het zijn dat er een koppelsleutel ontbreekt, om dan de data te kunnen koppelen, moet het mogelijk zijn om zelf een koppelsleutel te creëren. Wanneer koppelsleutels ontbreken, kan het gedrag wat is vastgelegd, niet gekoppeld worden aan de cliënt waar het om draait. Wanneer dit voor te veel databronnen het geval is, kan er geen individueel voorspelmodel gemaakt worden.

### 3.3.7 Mutatiegegevens

Mutatiegegevens moeten te herleiden zijn. Stel dat er in de tabel kinderen alleen de huidige status vermeldt staat, dan moet deze aangepast worden zodat er ook historische gegevens gebruikt kunnen worden. Hiervoor kan dan een mutatietabel voor kinderen bestaan waarin alle wijzigingen per cliënt zijn vastgelegd. Om te kunnen herleiden welke status bij welke cliënt hoort en bij welk kind, is er een mutatiecode nodig. Wanneer er in zo'n geval geen mutatiecode aanwezig is, heb je niks aan de gegevens omdat je dan geen historische gegevens hebt. Je kunt dan alleen de cliënten met elkaar vergelijken op het huidige moment. Dit geeft geen duidelijk beeld van de werkelijkheid. Om een goed model te maken, moet de hele tijdlijn van een uitkering in kaart zijn gebracht

### 3.3.8 Aantal bruikbare variabelen

Er moeten voldoende bruikbare variabelen aanwezig zijn om een voorspelmodel te kunnen maken. Stel dat er maar 3 variabelen gevuld zijn, met waar weinig variatie, dan is het onwaarschijnlijk dat daar een goede voorspelling uitkomt.

### 3.3.9 Aantal unieke cliëntnummers

Er zijn voldoende unieke cliëntnummers nodig om een goede voorspelling te kunnen maken. Zo kun je met maar 100 of 1000 cliëntnummers niet alle patronen terugvinden en minder goed generaliseren. Uit ervaring hanteren we ongeveer een minimum van 3000 cliëntnummers voor een robuust model.

### 3.3.10 Aantal gevallen van onrechtmatigheid

Er zijn voldoende gevallen nodig waarbij onrechtmatigheid is vastgesteld zodat het algoritme genoeg voorbeeld heeft om op te leren. Voor het voorspellen van onrechtmatigheid in de bijstand hanteren wij een minimum van 50 onrechtmatigheidsgevallen.

### 3.3.11 Privacy

De data moet gepseudonimiseerd zijn. Met pseudonimiseren worden persoonsgegevens getransformeerd in een dataset die niet meer direct herleidbaar is tot een persoon. De burgerservicenummers van personen mogen nooit gebruikt worden. Voor deze cliëntnummers zal de kans op onrechtmatigheden berekend worden. Wanneer gevoelige gegevens toch zijn aangeleverd, worden ze direct verwijderd. De gemeente wordt op de hoogte gesteld en een datalek wordt geregistreerd.

## 3.4 De gebruikte data

In de onderstaande tabel is elke kolom als databron weergegeven met in elke rij de variabelen die in deze bron voorkomen. Alle blauw gekleurde cellen geven de variabelen weer die uiteindelijk in het model zijn gebruikt. De andere cellen zijn variabelen die niet zijn meegenomen omdat ze niet voldoende kwaliteit met voldoende kwantiteit hebben. Daarnaast kunnen bepaalde gegevens ook niet zijn meegenomen omdat ze niet relevant zijn voor het voorspellen van onrechtmatigheid in de bijstand,

Tabel 1 - Variabelen per tabel:

| **Cliënten** | **Uitkeringsdossiers** | **Kinderen** | **Vakanties** | **Blokkades** | **Debiteuren** | **Blokkades inkomstenbriefjes** |
|---|---|---|---|---|---|---|
| Cliëntnummer | Cliëntnummer | Cliëntnummer | Cliëntnummer | Dossiernummer | Cliëntnummer | Cliëntnummer |
| Datum registratie | Dossiernummer | Volgnummer | Aanvraagnummer | Cliëntnummer | Debiteurnummer | Periodenummer |
| Geslacht | Cliëntnummer partner | Geboortedatum | Aanvraagdatum | Reden | Code groep | Dossiernummer |
| Geboortedatum | Datum registratie | Ten laste komend | Datum afdoening | Ingang blokkade | Soort vordering | |
| Overlijdensdatum | Aanvang uitkering | Woonsituatie | Code groep | Einde blokkade | Datum registratie | |
| Burgerlijke staat | Einde uitkering | Overlijdensdatum | Omschrijving | | Datum begin vordering | |
| Huisvesting | Regeling | Cliëntnummer kind | Code aard | | Datum eind vordering | |
| | Groep | Einddatum ten laste komend | Omschrijving | | Saldo | |
| | Aard | Geslacht | | | Sts vordering cbs | |
| | Soort uitkering | | | | Omschrijving vordering | |
| | Oorzaak uitkering cliënt | | | | Cliëntnummer partner | |
| | Oorzaak uitkering partner | | | | | |
| | Reden einde cliënt | | | | | |
| | Reden einde partner | | | | | |
| | Samenleving | | | | | |

| Rechten, plichten en ontheffingen | Participatietrajecten | Bijzondere situaties | Vermogen | Werkprocessen | Controles en fraude registraties |
|---|---|---|---|---|---|
| Cliëntnummer | Cliëntnummer | Cliëntnummer | Cliëntnummer | Cliëntnummer | Cliëntnummer |
| Code recht plicht | Soort traject | Code bijzondere situatie | Dossiernummer | Aanvraagnummer | Dossiernummer |
| Omschrijving | Trajectsoort | Bijzondere situatie | Peildatum | Aanvraagdatum | Datum besluit |
| Begindatum | Begindatum | Begindatum | Woning | Datum afdoening | Soort |
| Einddatum | Einddatum | Einddatum | Saldo | Code groep | Datum start fraude |
| | Code status | | Vrijlating | Aard verzoek | Datum eind fraude |
| | Status | | | Code aard | Datum start gedraging |
| | Activiteit | | | Soort werkproces | Datum eind gedraging |
| | Begindatum | | | Cliëntnummer partner | Categorie |
| | Einddatum | | | | Omschrijving |
| | | | | | Percentage/100 |
| | | | | | Bedrag boete / 100 |
| | | | | | Bedrag fraude / 100 |
| | | | | | Afw. Bedrag maatregel; |
| | | | | | Startdatum maatregel |
| | | | | | Einddatum maatregel |
| | | | | | Cliëntnummer partner |

# 4 Data verwerken

## 4.1 Definitie van onrechtmatigheden

Aan de hand van onrechtmatigheden uit het verleden leert het model herkennen wat mogelijke onrechtmatigheden in de toekomst zijn. Deze onrechtmatigheden zijn meestal vastgesteld na onderzoek van deze uitkeringsgerechtigde. In zo'n geval is er sprake van fraude in de wet werk en bijstand of in de bijzondere bijstand. Daarnaast valt een fraude terugvordering die geregistreerd staat in het debiteurenbestand ook onder de onrechtmatigheid definitie. De uitkering van de uitkeringsgerechtigde wordt na zo'n incident stopgezet, aangepast, voortgezet of blijft ongewijzigd.

## 4.2 Data preparatie en opschoning

Voordat een model getraind kan worden, moet de data geprepareerd zijn. Het preparen en opschonen van de data verloopt via de volgende stappen:

1. Brontabellen werkbaar maken;
2. Training en voorspel datasets maken en de peilmoment selectie;
3. Onrechtmatigheden definiëren in de data;
4. Basis dataframe en voorspel niveau;
5. Samenvoegen tabellen tot 1 dataset en creëren van variabelen;
6. Incomplete, dubbele of missende data corrigeren;

### 4.2.1 Brontabellen werkbaar maken

In deze stap laden we de data in en vertalen we de bron tabellen naar tabellen waarmee het model kan werken. Dit betekent dat bijvoorbeeld namen worden veranderd van tabellen en variabelen zodat er geen vreemde tekens meer in voorkomen, De modelleer technieken kunnen namelijk niet omgaan met vreemde tekens. Verder vertalen we variabelen en variabel categorieën naar standaard namen, zodat de betekenis van de variabelen meteen duidelijk is, maar ook zodat de scripts deels hergebruikt kunnen worden voor fraudemodellen van andere gemeentes.

### 4.2.2 Training en voorspel datasets maken en de peilmoment selectie

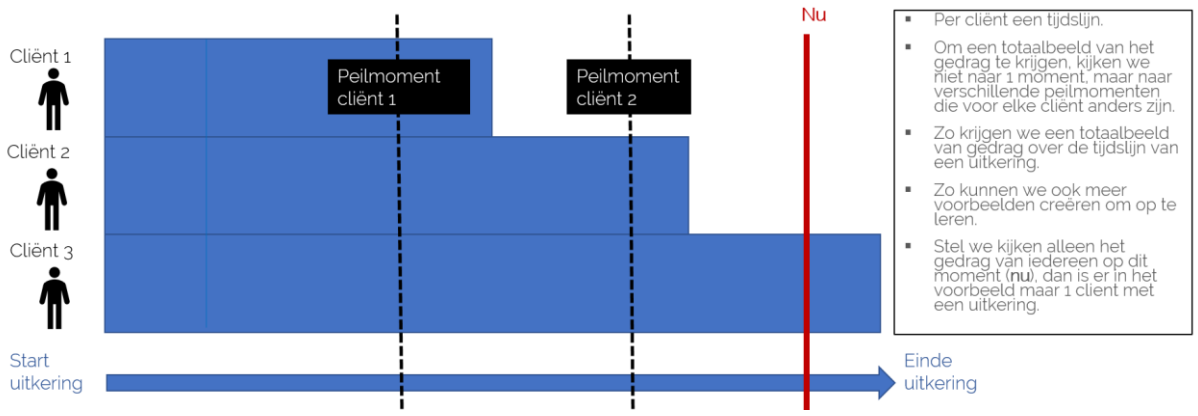
Om het model te leren welke patronen over het algemeen voorkomen bij onrechtmatigheden en welke niet, gaan we het model trainen. Het trainen van het model betekent eigenlijk dat we het model voorbeelden geven van onrechtmatig gedrag en dat het model gaat proberen om deze voorbeelden zelf in de data te vinden. Daarna testen we hoe goed het model instaat is om andere gevallen van fraude te vinden, die het model nog niet eerder heeft gezien. Wanneer we het model in de praktijk gaan gebruiken, willen we namelijk ook dat het model in staat is om nieuwe fraude gevallen te vinden.

Om dit te kunnen doen is de data opgesplitst in een dataset om op te trainen en een dataset om op te testen, de training dataset, en een dataset waarop de uiteindelijke voorspellingen zijn gedaan, de voorspel dataset. Eerst is train dataset gebruikt om het model op te trainen en op te testen totdat het uiteindelijke model is ontwikkeld. Daarna is de voorspel dataset gebruikt om de uiteindelijke voorspellingen op te doen In deze dataset is per cliënt een willekeurig moment binnen de uitkering gekozen waarop naar het gedrag van deze cliënt is gekeken. Dit moment is het peilmoment (Figuur 2)**.** Alle informatie uit andere bronnen voor elke cliënt is ook geselecteerd op dit peilmoment, dus informatie na het peilmoment wordt niet meegenomen. Door op deze willekeurige peilmomenten naar het gedrag van de cliënten te kijken, is ervoor gezorgd dat je op elk moment in de uitkering een goede voorspelling kan maken. Je zorgt namelijk dat op verschillende punten over de tijdslijn van de uitkering, gedrag in kaart is gebracht. Ook zorg je dat je de

informatie van alle personen in de uitkering kan gebruiken. Stel dat je alleen maar naar het gedrag zou kijken op 2018-01-01, dan mis je een heleboel uitkeringen die al zijn gestopt voor 2018-01-01 of die nog moeten beginnen. Je mist dus een heleboel voorbeelden waarop het model kan leren. Dus door met een peilmoment te werken, kun je de informatie van iedereen meenemen en op verschillende momenten over de tijdslijn van de uitkering. Op die manier krijgt het model zo'n realistisch mogelijk beeld van situatie en heeft het zoveel mogelijk situaties om op te leren.

Binnen de voorspel dataset is het peilmoment de huidige datum. Het is namelijk de bedoeling om op het huidige moment een voorspelling te doen om onrechtmatig gedrag op te sporen. Je hoeft dan ook alleen maar voor de mensen die op dit moment in de uitkering zitten een voorspelling te doen. Om die reden kun je wel de huidige datum als peilmoment gebruiken. Voor de uiteindelijk voorspelling gebruik je dan een model dat heeft geleerd op alle voorbeelden van alle uitkeringen uit de training dataset. Dat model gaat dan met alles wat het heeft geleerd opzoek naar onrechtmatig gedrag in de voorspel dataset, de mensen die op dit moment een uitkering hebben. Het model geeft dan alle personen uit de voorspel dataset een kans op onrechtmatigheid.

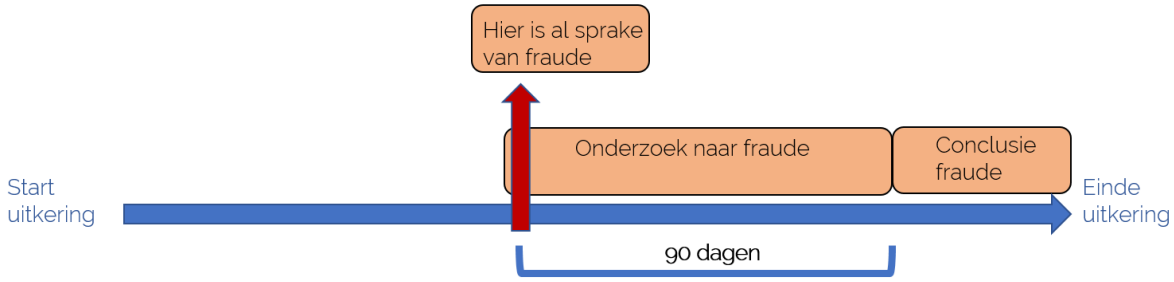
Figuur 2 - Weergave van het peilmoment:




### 4.2.3 Onrechtmatigheid definiëren in de data

De onrechtmatigheid is gedefinieerd in de data door een label van onrechtmatigheid te hangen aan alle cliënten die na een controle een conclusie van onrechtmatig gedrag in de wet werk en bijstand of in de bijzondere bijstand hebben gekregen, maar ook als er een fraude terugvordering is geweest. Vervolgens is het peilmoment van cliënten met onrechtmatig gedrag veranderd naar het moment 90 dagen voordat het onrechtmatigheidsonderzoek is gestart (Figuur 3). Bij gemeente Deventer zit er gemiddeld 90 dagen tijd tussen de start van het onderzoek en het concluderen van onrechtmatig gedrag. Dus het onrechtmatige gedrag vind al eerder plaats dan dat de conclusie is getrokken dat dit inderdaad onrechtmatig gedrag is. Door op dit moment te trainen, wordt het meest betrouwbare moment van onrechtmatig gedrag meegenomen in het model.

Figuur 3 - Weergave van tijdslijn onrechtmatig gedrag:




### 4.2.4 Basis dataframe en voorspel niveau

De basis dataframe is een tabel  met maar 1 uniek cliëntnummer per uitkering. Er is dus maar 1 regel per cliëntnummer (Tabel 2). Deze unieke cliëntnummers zijn nodig om een correcte koppeling te kunnen maken op het peilmoment. Door deze basisstructuur neer te zetten, is het mogelijk om een individuele voorspelling te maken, met genoeg onderscheidend vermogen voor het model, waarbij gelijk gecorrigeerd wordt voor eventuele tijdsinvloeden (wat vroeger was, hoeft nu niet meer zo te zijn), terwijl je wel wilt leren van de historie. Stel je zou de informatie niet samenvatten per persoon, maar je zou meerdere regels per persoon laten staan, dan zou het model minder goed kunnen voorspellen. Het model ziet namelijk elke regel als een voorbeeld. De voorbeelden die een samenvatting weergeven tot aan het peilmoment zijn veel informatiever dan de regels die alleen de informatie op het peilmoment weergeven. Tevens kan het dan voorkomen dat verschillende cases voor het model teveel op elkaar lijken, terwijl het model juist op zoek is naar het onderscheid, het verschil, van de verschillende cases.

Het is belangrijk dat we per persoon een voorspelling doen omdat de gemeente Deventer per persoon fraude onderzoeken uitvoert. De gemeente Deventer kan nu bijvoorbeeld de 10 personen met de hoogste kans op fraude selecteren en op deze personen een fraude onderzoek uitvoeren. Wanneer de gemeente willekeurig personen zou selecteren om te onderzoeken, vinden ze minder personen die daadwerkelijk fraude plegen. Door op deze manier te werken, kan de gemeente veel meer fraude opsporen dan voorheen. In het onderstaande figuur zie je een voorbeeld het voorspelniveau. Je ziet per cliëntnummer maar 1 regel met daarin informatie samengevat (zoals 'Aantal keer in de bijstand geweest').

Tabel 2 - Weergave van het voorspel niveau:

| Cliëntnummer | Peilmoment | Aantal keer in de bijstand geweest | Onrechtmatigheid |
|---|---|---|---|
| 1 | 2018-01-01 | 0 | 1 |
| 2 | 2018-10-01 | 0 | 0 |
| 3 | 2017-02-15 | 3 | 1 |

### 4.2.5 Samenvoegen tabellen tot 1 dataset & creëren van variabelen

Alle databronnen zijn gekoppeld aan de basis dataframe op cliënt niveau of dossier niveau en per databron zijn er variabelen gecreëerd. Per cliëntnummer is als het ware een samenvatting gegeven van het gedrag tot aan/op het peilmoment. Variabelen die bijvoorbeeld zijn gemaakt:

- Leeftijd; dit is de leeftijd op het peilmoment en die is berekend op basis van de geboortedatum.
- Aantal dossiers; dit is het aantal uitkeringsdossiers dat iemand heeft gehad tot aan het peilmoment.

### 4.2.6 Incomplete, dubbele of missende data corrigeren

Om met een modelleer techniek te kunnen werken, moeten alle missende gegevens gecorrigeerd zijn. Zo zijn voor alle numerieke waarden de missende gegevens vervangen met 0. Binnen categorische variabelen zijn de missende gegevens vervangen met 'geen'. Alle categorieën van categorische variabelen zijn meegenomen als aparte variabelen (dummies). Dit betekent dat iemand die in een bepaalde categorie valt, binnen de nieuwe 'dummy' variabele een 1-waarde krijgt en wanneer iemand erbuiten valt krijgt die een 0-waarde (Tabel 3). Voor elke categorie van de categorische variabele wordt er zo'n dummy variabele aangemaakt. Dit is gedaan omdat de modelleertechnieken niet kunnen werken met categorische waarden, maar wel met 1-en en 0-en.

Tabel 3 - Weergave van een categorische variabele en de dummy varianten ervan:

| Cliëntnummer | Leefvorm | Leefvorm Samenwonend | Leefvorm Samenwonend met kinderen | Leefvorm Alleen wonend |
|---|---|---|---|---|
| 1 | Samenwonend | 1 | 0 | 0 |
| 2 | Samenwonend met kinderen | 0 | 1 | 0 |
| 3 | Alleen wonend | 0 | 0 | 1 |

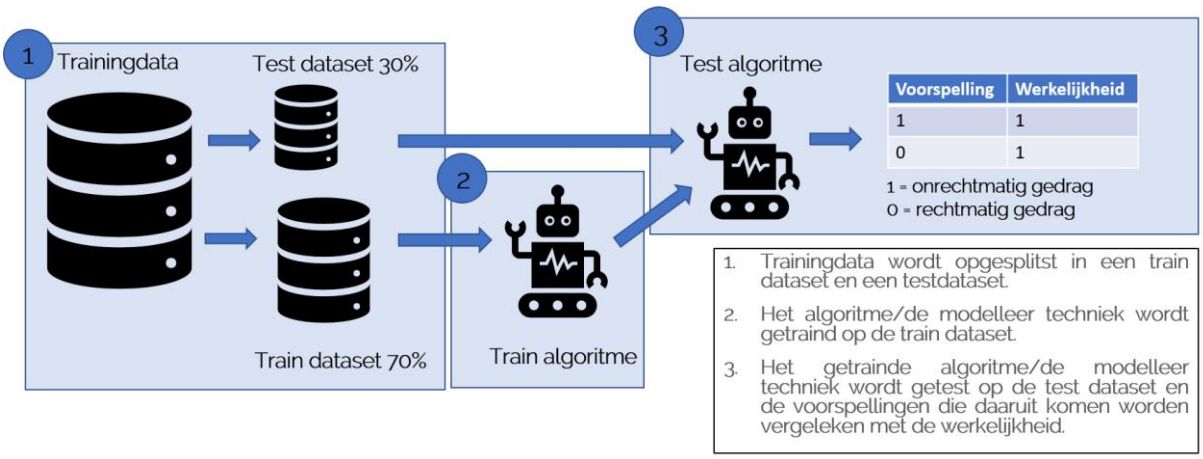
# 5 De modelontwikkeling

In de vorige hoofdstukken is besproken welke data verzameling en data verwerking nodig is voor het maken van een algoritme dat onrechtmatigheden binnen bijstandsuitkeringen voorspelt. In dit hoofdstuk wordt het keuzeproces beschreven voorafgaande aan het uiteindelijke model. Ook wordt uitgelegd welke modelleringstechnieken zijn gebruikt.

## 5.1 Aanpak modelkeuze

Om te bepalen welk model of samenstelling van modellen het beste werkt is de kwaliteit van de modellen getest. Dus eerst zijn de modellen getraind en daarna zijn getest op kwaliteit. Het trainen van het model betekent dat het model leert welke patronen over het algemeen voorkomen bij onrechtmatigheden en welke niet. We geven het model voorbeelden van onrechtmatig gedrag en dan gaat het model proberen om deze voorbeelden zelf in de data te vinden. Daarna testen we hoe goed het model instaat is om andere gevallen van fraude te vinden, die het model nog niet eerder heeft gezien. Wanneer we het model in de praktijk gaan gebruiken, willen we namelijk ook dat het model in staat is om nieuwe fraude gevallen te vinden. Om dit te kunnen doen is de trainingdata uit het vorige hoofdstuk is willekeurig verdeeld in een train dataset (70%) en een test dataset (30%). Met deze datasets zijn de modellen getraind en is er een voorspelling gemaakt (Figuur 4). Met train dataset leert het model fraude patronen herkennen en met de test dataset is de kwaliteit getest. Tijdens een test wordt er een voorspelling gedaan op de test dataset met het model dat was getraind op de train dataset. Deze voorspelling is vergeleken met de werkelijke gevallen van bijstandsonrechtmatigheden. Hierbij is gekeken naar de 10 gevallen met de hoogste voorspelde kans op onrechtmatigheid en vervolgens is vergeleken hoeveel gevallen daarvan daadwerkelijk onrechtmatig gedrag vertonen. Het model of combinatie van modellen met de hoogste verhouding correct voorspelde onrechtmatigheden in de top 10 is uiteindelijk geselecteerd en in de praktijk getest en gebruikt door Deventer.

Figuur 4 - Weergave van het trainen en testen van het algoritme:




## 5.2 Uitleg over toegepaste modellen en technieken

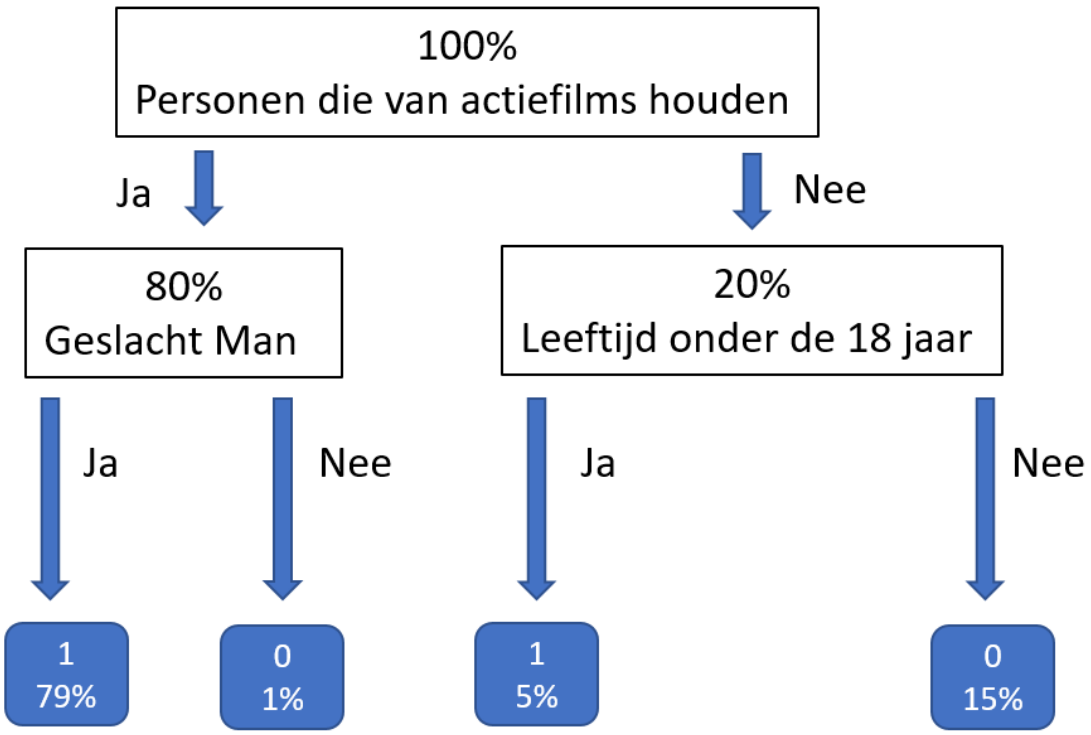
Voor het voorspellen van onrechtmatigheden bij bijstandsuitkeringen is de data getraind en daarna getest met verschillende technieken. Al deze technieken kunnen in de data patronen te herkennen en hiermee regels opstellen die een label toekennen aan nieuwe gevallen van onrechtmatig gedrag.

De onderstaande technieken zijn getest alvorens tot het uiteindelijke model te komen:

### 5.2.1 Random Forest

Random forest is een techniek waarbij het model op de data wordt getraind aan de hand van beslisbomen (Figuur 5). Dus het model leert op basis van voorbeelden van onrechtmatig gedrag om zelf onrechtmatig gedrag te herkennen. Een beslisboom is een overzicht van alle mogelijke uitkomsten van een reeks gerelateerde keuzes. Aan de hand van een beslisboom kunnen nieuwe gevallen geclassificeerd worden als onrechtmatig of rechtmatig. Om uit te leggen hoe een beslisboom eruit ziet, nemen we een voorbeeld waarbij we in kaart te brengen wie er naar de film 'Captain Marvel 'gaat. In het onderstaande voorbeeld zie je daar een visualisatie van:

Figuur 5 - Welke groepen mensen gaan er naar 'Captain Marvel'?

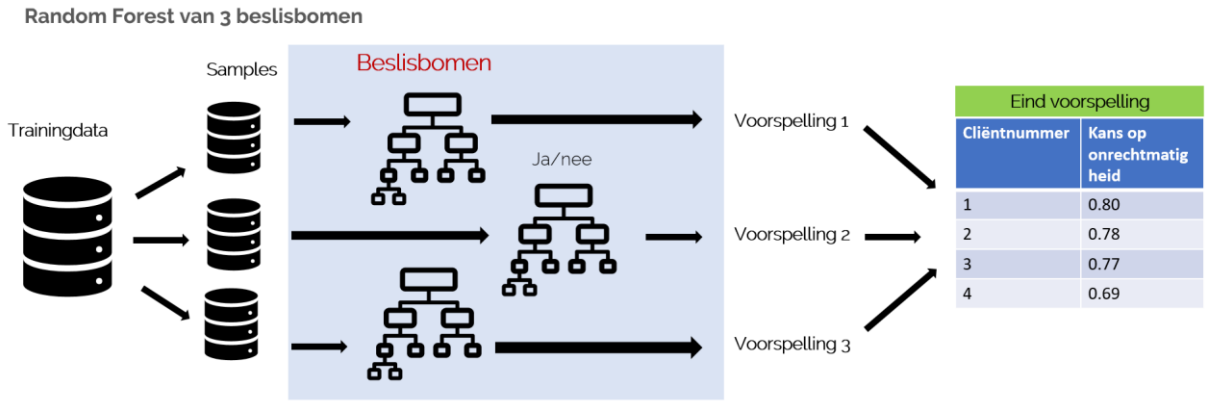



We zien hier dan de beslisboom 3 variabelen gebruikt om in te delen wie er wel en wie er naar 'Captain Marvel' gaat.: houdt een persoon van actiefilms, het geslacht en de leeftijd. De beslisboom heeft 2 groepen gevonden die naar 'Captain Marvel' gaan. In de eerste groep zit 79% van de personen uit de data en dit zijn allemaal mannen die van actiefilms houden. De andere groep bestaat uit personen die allemaal niet van actiefilms houden en jonger dan 18 jaar zijn. Deze groep bedraagt 5% van de personen in de dataset. Een beslisboom gebruikt dus een aantal variabelen om groepen te vinden die wel of niet in een bepaalde klasse vallen.

Wanneer de Random forest techniek gebruikt wordt voor het trainen van het model en voorspellen met het model, hebben we het over een 'bos' aan beslisbomen (Figuur 6). Het aantal beslisbomen dat in een Random forest model gebruikt wordt, kan gevarieerd worden in het model Meer bomen is altijd beter, omdat het naar een optimale verdeling convergeert. Echter, we gebruiken minder bomen, omdat dit simpelweg sneller is.. Door het aantal bomen in de tests te variëren vinden we het optimale model. Dit model

voorspeld zo veel mogelijk gevallen correct, maar is ook snel in het herkennen van patronen in de data. Het Random forest selecteert willekeurig een aantal variabelen en maakt met deze variabelen een beslisboom. Op deze manier kan er geen bias ontstaan waarbij bepaalde variabelen steeds geselecteerd worden om een patroon op te herkennen.. De kwaliteit van de verschillende beslisbomen wordt gemeten en geeft aan elke boom een gewicht, waarbij bomen die beter voorspellen een hoger gewicht krijgen dan minder goed voorspellende beslisbomen. Het uiteindelijk model zal nieuwe gevallen classificeren aan de hand van al deze beslisbomen en hun gewichten. Dit bos van beslisbomen is beter dan een enkele grote beslisboom omdat een grote beslisboom veel te specifiek zou worden voor een dataset. Die beslisboom zou dan niet meer goed werken op een nieuwe dataset. Door heel veel bomen te combinen, die op willekeurige variabelen zijn gebaseerd, kan er een algemeen patroon herkend worden. Dit patroon is dan ook te vinden in een nieuwe dataset.

Figuur 6 - Weergave van de Random forest techniek:




### 6.2.2 Neurale netwerken

Neurale netwerken zijn gebaseerd op de werking van het menselijk brein. Het brein bestaat uit 10 miljard neuronen die ieder verbonden zijn met 10,000 andere neuronen. Een neuron bestaat uit een cellichaam met daaraan dendrieten en een axon. De neuronen liggen in een enorm netwerk doordat de axonen van neuronen verbonden zijn met de dendrieten van andere neuronen. Een voorbeeld van de werking van het brein:

*Stel je loopt in het park en je ziet iets bewegen wat ongeveer 50 centimeter hoog is, een vacht en een staart heeft en je hoort het blaffen. Doordat je als kind hebt geleerd hoe een hond eruit ziet kan je vervolgens een classificatie geven: dit is een hond. Stel je loop nog een stukje verder door het park en je ziet alleen een stukje vacht. Is het van een hond, van een kat of van iets anders? Nu kun je niet classificeren van welk beest het stukje vacht is, omdat er niet voldoende signalen binnenkomen om te classificeren. Op neuronniveau worden er impulsen via je ogen en oren als input gegeven aan je brein, als elektronische signalen. Uiteindelijk komt via een neuron een signaal aan bij een synaps. Als gevolg van deze signalen komen kleine stofjes genaamd neurotransmitters vrij in de synaps spleet. Via de dendrieten van ons neuron worden deze neurotransmitters ontvangen. Wanneer er genoeg neurotransmitters worden losgemaakt en het ontvangen signaal boven een bepaalde grens uitkomt, wordt de neuron geactiveerd en zendt hij een signaal door de axon. Hierdoor gaat ons neuron zelf neurotransmitters afgeven in de synapsspleten verbonden met zijn axon. Belangrijk is dus dat er een signaal wordt gegeven als het inkomend signaal boven een bepaalde grens uitkomt. Deze drempel kan verschillen per inkomend signaal. Deze dynamiek tussen neuronen wordt gemodelleerd in Neurale netwerken.*

Neurale netwerken bestaan uit een input layer, 1 of meerdere hidden layers en een output layer (Figuur 7). De input layer bevat inputvariabelen en deze variabelen krijgen een willekeurig gewicht waarmee ze vermenigvuldigd worden en vervolgens opgeteld. Alle variabelen die we in het model willen gebruiken zitten in de input layer, dit zijn de meeste variabelen uit Tabel 1 - Variabelen per tabel. In het voorbeeld zijn dit er 3. De hidden layer bestaat uit een bepaald aantal neuronen, in het voorbeeld zien we 1 hidden layer met 4 neuronen. Het aantal neuronen is tijdens het testen van de techniek gevarieerd, net als het aantal hidden layers. De neuronen in de hidden layers zijn altijd een combinatie van alle inputvariabelen (de som van de gewogen variabelen). De neuronen in de hidden layer zorgen ervoor dat de inputwaarde (de som van de gewogen variabelen) zo getransformeerd wordt dat de uitkomstwaarde het dichts bij zijn werkelijke waarde zit. In dit geval is de uitgangswaarde een 1 (onrechtmatigheid) of een 0 (onrechtmatigheid). De output layer bestaat uit een combinatie van de neuronen uit de hidden layer, waarvan de waardes boven een bepaalde threshold uitkomen. Wanneer een neuron uit de hidden layer boven de threshold uitkomt, dan krijgt deze neuron een 1 en anders een 0 waarde. Van deze 0-en en 1-en wordt vervolgens weer een combinatie genomen met nieuwe willekeurige gewichten. Dit is de waarde van de output neuron in de output layer. Omdat we in het geval van het voorspellen van onrechtmatigheid in de bijstand de kans op 1 waarde voorspellen (het plegen van een onrechtmatigheid), is er sprake van 1 output neuron.

Tijdens het trainen van een neuraal netwerk model, leert het model wat de gewichten van de input variabelen moeten zijn om een voorspelling te doen die zo dicht mogelijk bij de werkelijkheid ligt. Stel we voorspellen per persoon de kans op fraude. Wanneer iemand fraude pleegt, is de fraude variabele 1 en wanneer er geen fraude is gepleegd, dan is de waarde 0. We voorspellen dus hoe groot de kans is dat iemand een 1 waarde heeft voor de fraude variabele. Stel we hebben maar 3 input variabelen zoals in het voorbeeld plaatje (Figuur 7).

**Het Neurale netwerk berekent de kans op fraude per persoon als volgt:**

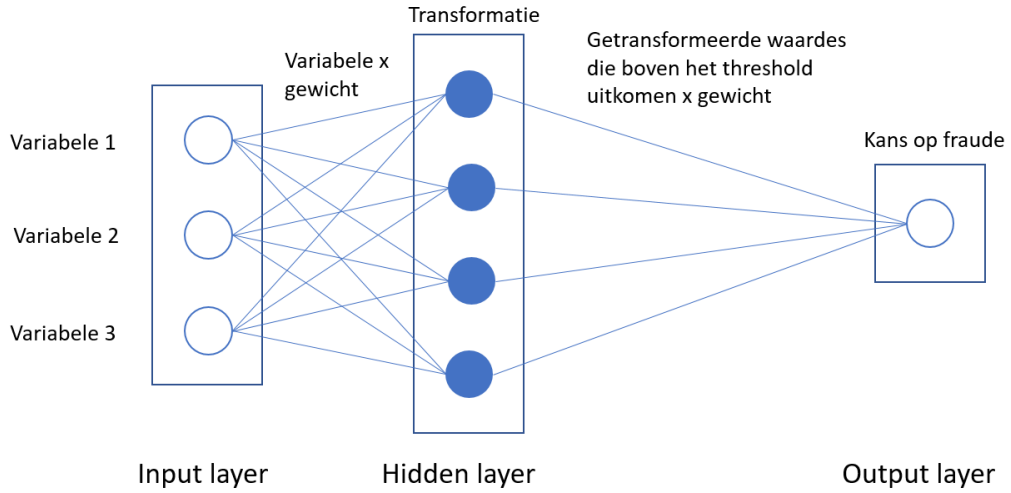
Fraude variabele = 1 als:

variabele 1 x gewicht 1 + variabele 2 x gewicht 2 + variabele 3 x gewicht 3 > threshold

Fraude variabele = 0 als:

variabele 1 x gewicht 1 + variabele 2 x gewicht 2 + variabele 3 x gewicht 3 <= threshold

Figuur 7 - Weergave van Neurale netwerken




### 6.2.3 Lasso

De lasso modeling techniek is een uitbreiding op de lineaire regressie analyse. Bij een regressie analyse wordt een afhankelijke variabele beschreven door een som van onafhankelijke variabelen vermenigvuldigd met een gewicht plus een restterm. Denk bijvoorbeeld aan de prijs van fruit, die is afhankelijk van het seizoen en van de markt. In dit geval hebben we het over onrechtmatig gedrag en dat is bijvoorbeeld afhankelijk van variabelen zoals leeftijd en aantal dossiers.. Bij een lineaire regressie analyse worden gewichten gezocht zodat het verschil tussen de voorspelling van onrechtmatigheid en de daadwerkelijke waarde zo klein mogelijk is. De gewichten geven aan in hoeverre een variabele invloed heeft op de uiteindelijke uitkomst.

Lasso modeling is een techniek die de voorspel kracht probeert te verbeteren door alleen de variabelen een gewicht toe te kennen die het model verbeteren. De overige variabelen zullen een gewicht van 0 krijgen en hierdoor niet gebruikt worden voor het classificeren van nieuwe gevallen. Op deze manier selecteert de techniek zelf de belangrijkste variabelen, waardoor je enigszins kunt corrigeren voor overfitting. Overfitting betekent dat het model te specifieke patronen heeft geleerd gedurende training, en daardoor niet goed voorspelt op nieuwe data. Uiteindelijk moet het model namelijk de algemene patronen herkennen die ook in de toekomstige data voorkomen. Dit zijn dus patronen die niet individueel zijn, maar die bij meerdere personen voorkomen. Gedragspatronen zijn immers nagenoeg niet gelijk. Verder wordt het model ook getest met verschillende lasso gewichten. We testen dit om dezelfde reden als het testen van het aantal bomen bij een random forest. Namelijk, het model moet optimaal presteren, de patroonherkenning mag niet te specifiek zijn, maar ook niet te onspecifiek.

# 5 Resultaten

Nu duidelijk is welke data en welke technieken zijn getest, bespreken we welke parameters zijn gebruikt om onrechtmatigheden in de bijstand te voorspellen.

## 5.1 Hoe werkt de modelbeoordeling

Om de kwaliteit van modellen te beoordelen worden de correcte en incorrecte voorspellingen met elkaar vergeleken. Dit is goed te vergelijken met ROC curves (Figuur 9). In een ROC curve worden de correcte voorspellingen afgezet tegen de incorrecte voorspellingen. De correcte voorspelling geeft aan dat het model onrechtmatigheid voorspelt waar dit ook daadwerkelijk het geval is. Bij de incorrecte voorspelling is ook onrechtmatigheid voorspeld, maar is dit in werkelijkheid niet het geval. De ROC curve is verder ook vergeleken met een curve van een model dat een willekeurige voorspelling geeft. Hoe groter de afstand tussen de ROC curve en de willekeurige curve, des te beter het model kan voorspellen. Daarbij is voornamelijk van belang dat er een groot verschil te zien is links, aan het begin, van de geplotte curves (zie de figuren hieronder). We willen namelijk vooral de gevallen die volgens het model de hoogste fraude kansen hebben, correct voorspellen.

De ROC curve is gebaseerd op de confusion matrix (Figuur 8). Naast de ROC curve kijken we ook naar de confusion matrix van de top 10 gevallen met de hoogste fraude kansen. Een confusion matrix geeft de volgende zaken weer:

- Daadwerkelijke aantal fraude gevallen
- Daadwerkelijk aantal gevallen zonder fraude
- Voorspelt aantal fraude gevallen
- Voorspelt aantal gevallen zonder fraude

Aan de hand van deze aantallen kan het percentage correct voorspelde fraude gevallen berekend worden. We willen uiteindelijk een model waarbij zoveel mogelijk gevallen binnen de top 10 hoogste fraude kansen correct voorspeld wordt. In de afbeelding hieronder zien we dat **3 van de 10 gevallen correct voorspeld** zijn. Dus 30% van de personen in de top 10 met de hoogste fraude kans is in dit voorbeeld correct voorspeld.

Figuur 8 – Confusion matrix

| | | Daadwerkelijke aantallen | |
|---|---|---|---|
| | | Wel fraude | Geen fraude |
| Voorspelde aantallen | Wel fraude | 3 | 2 |
| | Geen fraude | 3 | 2 |

*Deze afbeelding is een voorbeeld van een confusion matrix voor de top 10 hoogste fraude kansen. Je ziet dat bij elkaar opgeteld de getallen in de matrix 10 zijn. In totaal zijn er daadwerkelijk 6 fraude gevallen in de top 10 aanwezig. Er zijn 5 gevallen van fraude voorspeld in de top 10. Van het aantal voorspelde fraude gevallen zijn er 3 daadwerkelijk fraudeur en 2*

*zijn geen fraudeur. Het model voorspelt dat 3 van de daadwerkelijk fraude gevallen, geen fraude gevallen zijn.*

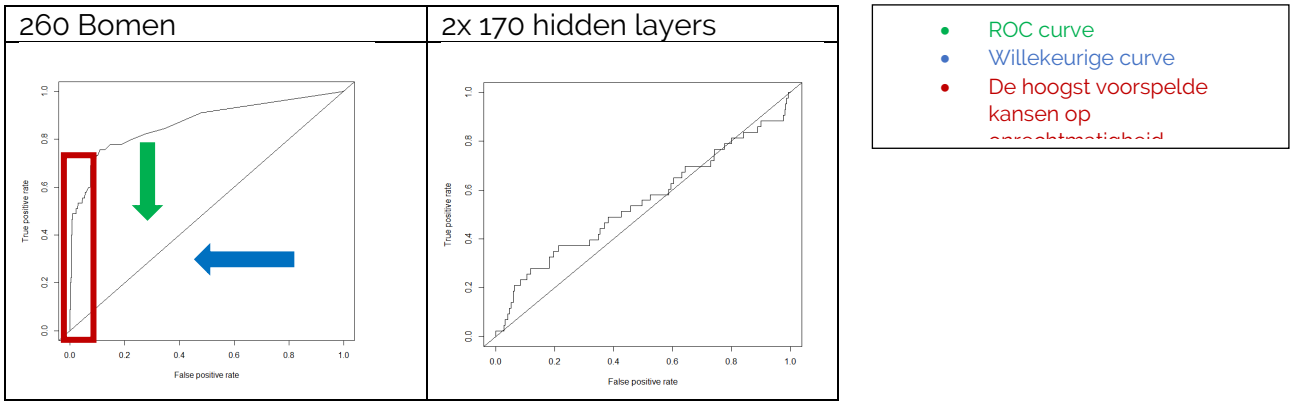
## 5.2 Modelleringsfase – assemblages, losse modellen

Een robuust model, of combinatie van modellen, zorgt ervoor dat er zowel bij het testen op bekende (in sample) als op onbekende (out of sample) data vergelijkbare resultaten uitkomen. Wanneer het algoritme is getest op bekende data, dan is het algoritme getest op dezelfde data als waarop het algoritme getraind is. Wanneer het model getest is op onbekende data dan is het algoritme getest op andere data dan waarop het getraind is. Het gaat hier dus om testdata die nog helemaal nieuw is voor het algoritme. Wanneer een model goed in sample kan voorspellen en minder goed out of sample kan voorspellen wordt dit 'overfitten' genoemd. Om ervoor te zorgen dat een model ook goed out of sample kan voorspelen wordt niet altijd het beste model gekozen. Een 'simpeler' model waar de kwaliteit van de voorspellingen bijna net zo goed zijn als het 'perfecte' model zal een betere voorspeller zijn voor nieuwe data. Door meerdere modellen te combineren kan overfitting worden tegengegaan. Op deze manier voorkom je dus een overfitting of underfitting in je voorspellingen.

Hieronder zien we per geteste modelleer techniek een voorbeeld van de testresultaten

**Random forest:** **Neuraal netwerk:**

Figuur 9 – ROC curves van Random forest en Neuraal netwerk




*Uitleg van de grafieken: Met het Random forest is er een test gedraaid met alle variabelen waarbij we nu een voorbeeld laten zien waarin het aantal beslisbomen op 260 bomen is gezet. Bij het neurale network zijn 2 hidden layers van 170 neuronen gebruikt. Er is te zien dat de ROC curve van een Random forest met 260 bomen het beste resultaat weergeeft vergeleken met het neurale netwerk. Het verschil tussen de ROC curve en het willekeurige model is namelijk voor het random forest een stuk groter dan bij het neurale netwerk.*

Voor het tegengaan van overfitten is het combineren van verschillende modellen een goed hulp middel. Om die reden is er gekozen om het Random forest en de neurale netwerken te combineren tot een uiteindelijk model. Omdat het Random forest veel beter presteert krijgt dit model wel een groter gewicht dan het neurale netwerk.Het model maakt voor 70% gebruik van de voorspellingen van het Random forest model (met 260 bomen) en voor 30% van het Neurale netwerk model (met 2 hidden layers van 170 neuronen).

## 6. Proces van voorspellen en leren

Bekend is welke data gebruikt wordt, hoe deze geprepareerd is en welk technieken gebruikt zijn voor het vinden van een model dat zo goed mogelijke voorspellingen kan maken. De volgende stap is het maken van voorspellingen voor personen die op dit moment in de bijstand zitten. Om deze voorspelling te maken zijn per oplevering de volgende stappen uitgevoerd:

- De meest recente data van bijstandsgerechtigden is aangeleverd via een encrypted usb stick aan Deventer.
- Deze data bevat ook de bijstandsgerechtigden die bij een eerdere oplevering zijn onderzocht door de gemeente (de top 10). Voor deze personen is het dan bekend of ze wel of niet fraude hebben gepleegd.
- Het model wordt vervolgens getraind op de data, inclusief de nieuwe gevonden fraude cases. Op deze manier leert het model nog beter fraude voorspellen.
- Met dit nieuwe getrainde model is een voorspelling gemaakt op basis van de voorspel dataset oftewel de nieuw aangeleverde data, waarin alleen cliënten voorkomen die momenteel een uitkering hebben.
- De voorspellingen zijn vervolgens aangeleverd aan Deventer via de encrypted usb stick verbinding.
- De gemeente Deventer gaat vervolgens de bijstandsgerechtigden uit de top 10 onderzoeken op fraude.
- De opleveringen worden gemiddeld een keer per kwartaal gedaan.

Voor elke lopende bijstandsgerechtigden is een kans berekend voor mogelijke onrechtmatigheden. Deze zijn van hoog naar laag gesorteerd en zijn samen met de gepseudonimiseerde cliëntnummers met de gemeente Deventer gedeeld om nader onderzocht te worden.