Gemeente Rotterdam

Uitleg Risico-inschattingsmodel

Model gebruikt voor de selectie van werkzoekenden voor heronderzoeken

Opdrachtgever	W&I Beheer inkomen
Datum	24 augustus 2021

Inhoudsopgave

Hoofdstuk 0: Aanleiding....3
Hoofdstuk 1: Definitie algoritme en soorten doelstellingen....5
Hoofdstuk 2: De gebruikte gegevens....8
Hoofdstuk 3: Trainen van het algoritme....11
Hoofdstuk 4: De output....17
BIJLAGE: Uitleg variabelen en features....18

Hoofdstuk 0: Aanleiding

Cluster Werk & Inkomen (W&I) van de gemeente Rotterdam voert jaarlijks, op verzoek van de gemeenteraad, heronderzoeken uit. Deze onderzoeken hebben tot doel om de rechtmatigheid van de verstrekte uitkering te kunnen beoordelen en ook om te bezien of de uitkeringsgerechtigde behoefte heeft aan ondersteuning (bijvoorbeeld op gebied van armoede en schulden).

Van de voor enig jaar geplande heronderzoeken wordt sinds 2018 een gedeelte geselecteerd op basis van de uitkomsten van een risico-inschattingsmodel. De onderzoekstijd die beschikbaar is voor het doen van heronderzoeken is namelijk beperkt. Door gebruik te maken van het risico-inschattingsmodel kan die tijd gericht worden ingezet op het beoordelen van de rechtmatigheid in die gevallen, waarin de kans op onrechtmatigheid het grootst is. Zo kunnen eventuele onrechtmatigheden snel(ler) worden opgespoord en worden omvangrijke terugvorderingen in de toekomst zo veel mogelijk voorkomen. Niet alleen de gemeente maar óók de uitkeringsgerechtigde heeft hier baat bij. Daarnaast is de meerwaarde van het risico-inschattingsmodel, dat op basis van dit model alleen die uitkeringsgerechtigden voor een heronderzoek zullen worden uitgenodigd waarbij de kans op onrechtmatigheid relatief hoog worden ingeschat. Wordt de kans op onrechtmatigheid relatief laag ingeschat, dan zal de uitkeringsgerechtigde ook niet op basis van de uitkomsten van het risico-inschattingsmodel worden uitgenodigd voor een heronderzoek.

Het door W&I gebruikte risico-inschattingsmodel is gebaseerd op een algoritme. Het gebruik van dit algoritme is onderdeel geweest van het recente onderzoek van de Rekenkamer Rotterdam naar het gebruik van algoritmen door de gemeente Rotterdam[1].
De Rekenkamer heeft in haar rapport een aantal aanbevelingen gedaan, waaronder het vergroten van de uitlegbaarheid van dit algoritme. Ook Concern Auditing[2] heeft in haar advies gesteld dat er op dit punt belangrijke stappen moeten worden gezet.
Dat is het doel van deze notitie: uitleg verschaffen over de ontwikkeling en de werking van het algoritme dat in het risico-inschattingsmodel wordt gebruikt.

Deze notitie is bedoeld om uitleg te verschaffen over de ontwikkeling en het gebruik van het algoritme dat als basis dient voor het risico-inschattingsmodel.

Naast deze notitie zijn de volgende stukken relevant:
- ‘Heronderzoeken Werk en Inkomen (Achtergrond, werkwijze en systematiek heronderzoeken)’ met daarin een basale uitleg van de heronderzoekssystematiek;
- Het communicatieplan met betrekking tot de wijze waarop met de Rotterdammer wordt gecommuniceerd over het bestaan, het gebruik en de werking van het algoritme; en
- De procesbeschrijving, waarin alle stappen uitgewerkt zijn van het beschikbaar komen van de data die in het algoritme worden gebruikt, tot aan de risico-inschatting en de filtering die daarop plaatsvindt.

Zowel het communicatieplan als de procesbeschrijving worden voor 1 september 2021 afgerond.

In deze notitie wordt achtereenvolgens ingegaan op de volgende onderwerpen:
- Definitie van het begrip algoritme en duiding van de belangrijkste soorten doelstellingen bij het gebruik van algoritmen (hoofdstuk 1);

[1] Het rapport Gekleurde technologie is de downloaden via deze link.
[2] ‘Advies Analytics Uitkeringsonrechtmatigheid’ d.d. 4 februari 2021.

- De voor het algoritme gebruikte gegevens (hoofdstuk 2);
- Training van het algoritme (hoofdstuk 3);
- De output die het algoritme oplevert (hoofdstuk 4).

Elk hoofdstuk eindigt met een korte samenvatting.

In de bijlage is een beschrijving opgenomen van hoe de in het model gebruikte ‘features’ tot stand komen.

Hoofdstuk 1: Definitie algoritme en soorten doelstellingen

Wat verstaan we onder het begrip 'algoritme'?
Voor de definitie van wat onder de term 'algoritme' wordt verstaan, wordt aangesloten bij de definitie in de 'Richtlijnen algoritmen'[3] van het ministerie van Justitie en Veiligheid:

> "Een wiskundige formule of model, dat wordt uitgevoerd middels een computer, om een probleem op te lossen, een vraag te beantwoorden, een voorspelling te doen, een beslissing te nemen of die te ondersteunen."

Met andere woorden, het gaat hier *niet* om een door de mens gehanteerd beslisschema. Dat biedt ook duidelijke voordelen: omdat we voor het ontwikkelen en gebruik van het model door een computer berekeningen laten uitvoeren, kunnen we gebruik maken van veel een groot aantal combinaties van variabelen die efficiënt kunnen worden doorgerekend. Voor een dergelijk model kan wel voldoende rekenkracht worden geleverd door een zware computer, terwijl het voor een mens een vrijwel onbegonnen klus zou opleveren om een dergelijke grote hoeveelheid rekenregels door te rekenen.

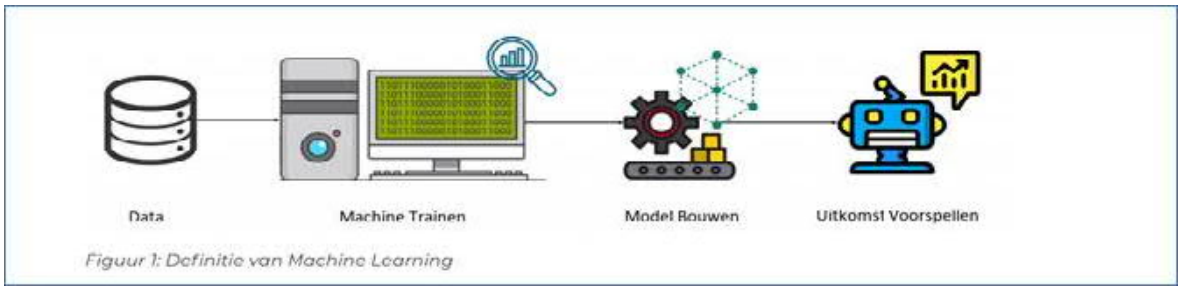
Een algoritme kan voor een buitenstaander al snel als 'eng' of 'bedreigend' overkomen, omdat er sprake lijkt van een black box: als buitenstaander weten we niet precies wat er 'in het model' gebeurt. Van belang is te beseffen, dat het niet zozeer een mysterie of geheim is wat er in het model gebeurt (dat gebeurt op basis van wetenschappelijk erkende methodieken, zie hoofdstuk 3, en kan dus ook worden nagerekend of gereproduceerd). Het probleem is alleen dat wat er in het algoritme gebeurt *zo veel* en *intensief* rekenwerk vergt, dat het door een mens lastig te reproduceren is.

Welk doel dient het algoritme?
In de eerder aangehaalde 'Richtlijn algoritmen', wordt ook uiteengezet voor welke doeleinden een algoritme kan worden gebruikt. Deze indeling varieert van beschrijvend (het algoritme beschrijft slechts wat er wordt waargenomen) tot voorschrijvend (het algoritme schrijft voor welk besluit er moet worden genomen).
Het algoritme dat wordt gebruikt voor het risico-inschattingsmodel, is een voorspellend algoritme. Op basis van het algoritme wordt een voorspelling gedaan over de kans op *onrechtmatigheid* of juist *rechtmatigheid* bij een groep uitkeringsgerechtigden. Hoe die inschatting door het algoritme wordt gemaakt, wordt uiteengezet in hoofdstuk 3.

> Het algoritme dat gebruikt wordt voor het risico-inschattingsmodel, is een voorspellend algoritme.

In de figuur op de volgende pagina is schematisch weergegeven hoe een voorspellend algoritme werkt.

[3] Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses d.d. 1 maart 2021. De richtlijnen kunnen via deze link worden gedownload.



Figuur 1: illustratie voorspellend algoritme[4]

Van belang is te beseffen dat het algoritme een inschatting maakt van de kans op (on)rechtmatigheid per individuele werkzoekende. Het gaat om een voorspelling over de *kans* op onrechtmatigheid (of juist rechtmatigheid) en dus geen *zekerheid*.
Het is op voorhand duidelijk dat het model niet met 100% zekerheid kan voorspellen bij welke werkzoekende sprake zal zijn van onrechtmatigheid, of juist rechtmatigheid. Daarvoor is de beoordeling of een uitkering op grond van de PW correct wordt verstrekt, ook veel te complex. Die beoordeling is van te veel factoren afhankelijk om door een computer te worden uitgevoerd en vergt bovendien maatwerk. Hierbij is dus *altijd* sprake van een inhoudelijke beoordeling door een medewerker (inkomensconsulent). Het algoritme neemt geen besluiten (al dan niet zelfstandig) en schrijft ook geen besluiten voor.

Voor de volledigheid: Het heronderzoek voor een werkzoekende die op basis van het risico-inschattingsmodel is geselecteerd, is ook inhoudelijk gelijk aan het heronderzoek bij een werkzoekende die op andere wijze (bijvoorbeeld a-select) is geselecteerd. Dezelfde stukken worden uitgevraagd, het gesprek vindt op dezelfde wijze plaats en het onderzoek leidt onder gelijke omstandigheden tot gelijke conclusies. Van vooringenomenheid in de zin dat het heronderzoek anders zou verlopen bij uitkeringsgerechtigden die op basis van het risico-inschattingsmodel worden geselecteerd, is dus géén sprake.

De inhoudelijke beoordeling van de rechtmatigheid wordt altijd door een mens gedaan. De computer neemt daarover geen beslissingen. De inhoud van het heronderzoek is bij alle uitkeringsgerechtigden gelijk.

Daarnaast wordt benadrukt dat onrechtmatigheid nog wat anders is dan fraude. Onrechtmatigheid betekent dat de verstrekte uitkering niet aansluit op de situatie van de uitkeringsgerechtigde en zijn eventuele gezin.
Dat *kan* in sommige gevallen het gevolg zijn van het verstrekken van onjuiste of onvolledige informatie aan gemeente[5]; en in een klein deel van dergelijke gevallen zal er sprake zijn van het bewust/opzettelijk (dus verwijtbaar) verstrekken van onjuiste of onvolledige informatie die leidt tot het ten onrechte ontvangen van bijstand of te veel bijstand (fraude).

Maar vaak gaat het om het onbewust verstrekken van onjuiste / onvolledige informatie (omdat het de burger niet duidelijk was dat een gegeven relevant zou zijn voor zijn recht op bijstand); of zijn meldingen van relevante wijzigingen bij het verkeerde 'loket' gedaan. Het komt zelfs voor dat wijzigingen wel zijn doorgegeven, maar niet correct zijn verwerkt door de gemeente.
Om in deze gevallen vervelende consequenties te voorkomen (zoals een omvangrijke terugvordering), is het van belang om rechtmatigheidsonderzoeken te blijven uitvoeren.

[4] Bron: Machine learning: uitleg en voorbeelden van Machine Learning - Divtag
[5] Op grond van artikel 17 lid 1 PW is de werkzoekende verplicht om 'op verzoek of onverwijld uit eigen beweging mededeling van alle feiten en omstandigheden waarvan hem redelijkerwijs duidelijk moet zijn dat zij van invloed kunnen zijn op zijn arbeidsinschakeling of het recht op bijstand.'

Vooral in die gevallen waarin de kans op onrechtmatigheid hoger wordt geschat dan in andere gevallen.
Dit betekent andersom geredeneerd, dat uitkeringsgerechtigden waarbij de kans op onrechtmatigheid relatief *klein* wordt geacht, minder kans lopen om te worden uitgenodigd voor een heronderzoek op basis van dit risico-inschattingsmodel[6]. Dit is een belangrijk voordeel van het gebruik van het risico-inschattingsmodel.

Het is heronderzoek is ontstaan en primair gericht op een beoordeling van de rechtmatigheid en dus niet per sé gericht op het opsporen van fraude.

Samenvattend:

- Het voor het risico-inschattingsmodel gebruikte algoritme is een voorspellend en dus géén voorschrijvend algoritme. Het inhoudelijke onderzoek naar het recht op uitkering wordt altijd door een mens (inkomensconsulent) verricht.
- Het heronderzoek dat wordt uitgevoerd bij een uitkeringsgerechtigde die op basis van het risico-inschattingsmodel wordt geselecteerd, vindt daarbij op dezelfde wijze plaats als een heronderzoek bij een uitkeringsgerechtigde die op andere wijze is geselecteerd.

[6] Zoals eerder aangegeven worden er ook jaarlijks heronderzoeken op andere wijze geselecteerd, waaronder a-select. Een lage risico-inschattingsscore betekent dus niet dat er een garantie is dat de werkzoekende in enig jaar *niet* wordt uitgenodigd voor een heronderzoek; alleen de kans dat de uitnodiging uitgaat op basis van het risico-inschattingsmodel is kleiner.

Hoofdstuk 2: De gebruikte gegevens

Welke gegevens worden als uitgangspunt gebruikt om het algoritme te ontwikkelen?
W&I heeft al langere tijd ervaring met het beoordelen van de rechtmatigheid van verstrekte bijstand. Vanuit die historie hebben we informatie verzameld over uitkeringsgerechtigden waarbij:

- voorheen een onrechtmatigheid is geconstateerd[7] (waarbij onrechtmatigheid is: de conclusie dat er voorheen te veel of ten onrechte uitkering is verstrekt);

of juist

- de conclusie is getrokken dat de uitkering over de voorliggende periode volledig correct is verstrekt.

Benadrukt wordt dat informatie over beide conclusies (onrechtmatigheid of juist rechtmatigheid), van belang is om het algoritme te kunnen 'trainen'. Het algoritme moet immers kunnen voorspellen of de kans op *onrechtmatigheid* relatief groter is; of juist de kans op *rechtmatigheid* relatief groter is (zodat de kans op onrechtmatigheid relatief klein is).

Naarmate W&I meer rechtmatigheidsonderzoeken heeft afgerond, beschikken we over steeds meer informatie; de conclusies van rechtmatigheidsonderzoeken die in 2020 zijn uitgevoerd, wegen weer mee in de ontwikkeling van het model voor 2021 en volgende jaren.

Om het algoritme te kunnen bouwen, worden gegevens van eerder uitgevoerde rechtmatigheidsonderzoeken gebruikt.

Welke variabelen worden gebruikt?
Per uitkeringsgerechtigde waaraan een uitkering is verstrekt, wordt een grote hoeveelheid gegevens in onze administratie vastgelegd. Het gaat dan om gegevens die W&I nodig heeft voor de uitvoering van haar wettelijke taken in het kader van het verstrekken van uitkering.

Deze gegevens worden vastgelegd in twee systemen:

- Socrates: Dit is de uitkeringsadministratie, waarin de hoogte van de verstrekte uitkering wordt vastgesteld en vastgelegd; en

- RMW/Raak: Dit is het systeem waarin de gegevens over de re-integratie van de uitkeringsgerechtigde worden vastgelegd. Het gaat dus om gegevens over de competenties en kwaliteiten van de uitkeringsgerechtigde die van belang zijn voor diens kansen op de arbeidsmarkt; en de trajecten die worden uitgezet om deze kansen te vergroten.

Door bij uitkeringsgerechtigden waarbij eerder een rechtmatigheidsonderzoek is uitgevoerd te gaan kijken naar de gegevens die in de administratie zijn vastgelegd, kan *met hulp van de computer* worden onderzocht of er een verband kan worden gevonden tussen:

- de gegevens zoals vastgelegd in de administratie enerzijds;
- en de eindconclusie van een onderzoek naar de rechtmatigheid (uitkering is onrechtmatig of rechtmatig verstrekt) anderzijds.

Benadrukt wordt dat er dus alleen gegevens worden gebruikt die beschikbaar zijn in de *eigen administratie*; en die gegevens zijn *noodzakelijk voor de uitvoering van de wettelijke taak* die W&I uitvoert. Er worden géén gegevens gebruikt uit andere administraties.

[7] De conclusie over het rechtmatig of onrechtmatig verstrekken van bijstand is altijd door een mens getrokken en vastgelegd.

De gebruikte variabelen bestaan uit gegevens die in het kader van de wettelijke taak zijn vastgelegd in de eigen administratie.

Hoeveel variabelen worden als uitgangspunt gebruikt?
In de administratie (RMW/Raak en Socrates) staan veel verschillende velden ('variabelen') welke soms ook voor ingevulde keuzes bevatten (de medewerker moet dan kiezen uit bepaalde 'waarden'). Dit levert een lijst op van 864 mogelijk te gebruiken waarden of variabelen van waarden. In de bijlage is nadere uitleg opgenomen over de gebruikte variabelen.

Worden al deze gegevens in het model gebruikt?
Nee, er wordt aan de voorkant al een filtering toegepast. Niet alle data zijn even goed bruikbaar om als input te gebruiken. Er zijn verschillende redenen om een waarde of variabele te verwijderen, bijvoorbeeld:

- Incompleetheid
  Als gegevens niet consequent zijn (in)gevuld, maakt dat zo'n gegeven ook minder goed bruikbaar binnen het algoritme. Wil een variabele bij kunnen dragen aan het doen van een voorspelling, dan is het van belang dat de variabele in voldoende gevallen gevuld is. Dat hangt onder meer af van de vraag of het systeem invulling afdwingt; maar in veel gevallen ook van de invuldiscipline bij de medewerker die verantwoordelijk is voor het invullen of aanvullen van gegevens in de administratie.

- Identieke invulling
  Als een variabele (bijna) altijd dezelfde waarde heeft is deze uiteraard niet bruikbaar om onderscheid te maken in het model. Zo'n variabele kan dus ook niet bijdragen aan het doen van een voorspelling en wordt dus ook niet in het model gebruikt.

- Samenhang tussen variabelen
  Als meerdere variabelen per persoon sterk samenhangen, heeft het geen toegevoegde waarde om ze beiden/allemaal mee te nemen. Één van die variabelen volstaat dan.
  Stel bijvoorbeeld dat een busbedrijf inzicht wilt krijgen in de grootte van de stoelen die moeten worden geplaatst. Het bedrijf beschikt in dit fictieve voorbeeld over uitgebreide informatie over eerdere passagiers, waaronder lengte en schoenmaat. Omdat een langer persoon doorgaans ook grotere voeten heeft en er dus een sterke samenhang is tussen beiden, heeft het géén zin om beide variabelen mee te nemen in het beslissingsproces.

Ook om overige inhoudelijke redenen, zoals bruikbaarheid/betrouwbaarheid van data, kan verdere filtering plaatsvinden.

Discriminatoire variabelen zoals nationaliteit worden (voor zover die al in de administratie worden vastgelegd) uiteraard niet in het model meegewogen[8].

Niet alle beschikbare gegevens worden gebruikt: bijvoorbeeld incomplete variabelen of identieke variabelen worden uitgefilterd. Discriminatoire variabelen op basis van nationaliteit worden niet in het model opgenomen.

[8] In het model dat in 2020 en 2021 is gebruikt, waren de variabelen leeftijd en geslacht wel opgenomen. Deze variabelen worden in volgende jaren buiten het model gelaten.

Is er een lijst van gebruikte variabelen beschikbaar?
Na de hierboven bedoelde schifting resteren er ca. 320 variabelen die worden gebruikt om het algoritme te 'trainen'. Hoe dat trainen in zijn werk gaat, wordt in het volgende hoofdstuk uiteengezet.

Wel wordt hierbij benadrukt dat de lijst van variabelen per jaar kan verschillen. Het algoritme wordt jaarlijks opnieuw getraind en de lijst van variabelen wordt jaarlijks gescreend, zodat er voor een volgend jaar sprake kan zijn van een andere lijst van variabelen die feitelijk binnen het model zijn toegepast.

Samenvattend:

- Voor de ontwikkeling van het algoritme worden historische gegevens gebruikt.
- Deze gegevens komen uit de eigen administratie. Het zijn gegevens die in het kader van de wettelijke taak worden vastgelegd. Er worden géén gegevens gebruikt uit andere administraties.
- Niet alle gegevens uit de administratie worden gebruikt. Incompleetheid, identieke invulling of samenhang tussen variabelen kan leiden tot het buiten beschouwing laten van bepaalde variabelen.
- Discriminatoire variabelen op basis van nationaliteit zijn niet opgenomen in het model.

Hoofdstuk 3: Trainen van het algoritme

Wat betekent het trainen van het algoritme?
Doel van het risico-inschattingsmodel is, zoals eerder aangegeven, een inschatting te maken (voorspelling) van de kans op onrechtmatigheid of juist rechtmatigheid voor individuele uitkeringsgerechtigden.. Daarbij wordt gebruik gemaakt van gegevens uit eerdere rechtmatigheidsonderzoeken, waaronder eerder uitgevoerde heronderzoeken.

Het algoritme zoekt naar verbanden tussen de gegevens die van uitkeringsgerechtigden bekend zijn uit eerdere rechtmatigheidsonderzoeken. Via een diepgaande analyse (zie hierna) onderzoekt 'de computer' of er in deze historische data een samenhang is te vinden tussen:

- De gegevens die in de administratie aanwezig zijn met betrekking tot uitkeringsgerechtigden waarbij de rechtmatigheid is onderzocht; en

- De uiteindelijke uitkomst van het onderzoek (onrechtmatige of juist rechtmatige verstrekking van uitkering).

Dit doet de computer uiteraard niet zelfstandig: hij wordt daarbij geprogrammeerd en geïnstrueerd door ter zake deskundige medewerkers (data scientists).
Het algoritme moet 'getraind' worden (moet leren) om die verbanden te kunnen vinden en een steeds betere voorspelling te doen. Gezocht wordt naar die combinatie van gegevens in de administratie, waarbij het grootste verband aanwezig is met onrechtmatige, of juist rechtmatige verstrekking. Dit vergt tijd, waarbij de ontwikkelaar kritisch kijkt naar tussentijdse uitkomsten en het algoritme 'begeleidt' op weg naar een optimaal model.

Benadrukt wordt dat dit trainen *jaarlijks opnieuw* plaatsvindt. Er is dus niet sprake van een algoritme dat 'af' is in de zin dat het algoritme het volgende jaar exact hetzelfde blijft. Het algoritme wordt jaarlijks aangepast op basis van de op dat moment beschikbare data.
Hier zouden we de metafoor kunnen gebruiken van een voetbalelftal, dat bij elke voorbereiding voor een nieuw toernooi of kwalificatie bestaat uit (deels) andere voetballers met verschillende kwaliteiten [de variabelen], die daardoor ook op andere wijze worden getraind en uiteindelijk opgesteld [het model].

Hoe werkt het trainen van het algoritme?
Bij het trainen van het algoritme kan gebruik worden gemaakt van verschillende modelleertechnieken. Bekende voorbeelden zijn Random forest, Gradient Boosting en Logistische regressie. Ook de uiteindelijk gebruikte methode kan van jaar tot jaar verschillen, omdat er dan opnieuw wordt gezocht naar de methode die het optimale resultaat oplevert.

Al deze methodieken:
- gaan uit van wiskundige theorie;
- zijn wetenschappelijk onderbouwd;
- zijn internationaal erkend; *maar*
- laten zich moeilijk uitleggen voor de lezer.

Dat laatste heeft in het bestek van deze notitie ook weinig toegevoegde waarde omdat de toegepaste methodiek, zoals aangegeven, *jaarlijks* kan verschillen. Mocht de lezer zich daar toch meer in willen verdiepen, dan is een voorbeeld van een naslagwerk dat hierover kan worden geraadpleegd de uitgave 'Introduction to Statistical Learning (G. James e.a.)', dat ook als pdf kan worden gedownload; of het boek Artificial Intelligence. A modern Approach (Third Edition) van Stuart Russell en Peter Norvig dat eveneens als pdf kan worden gedownload.

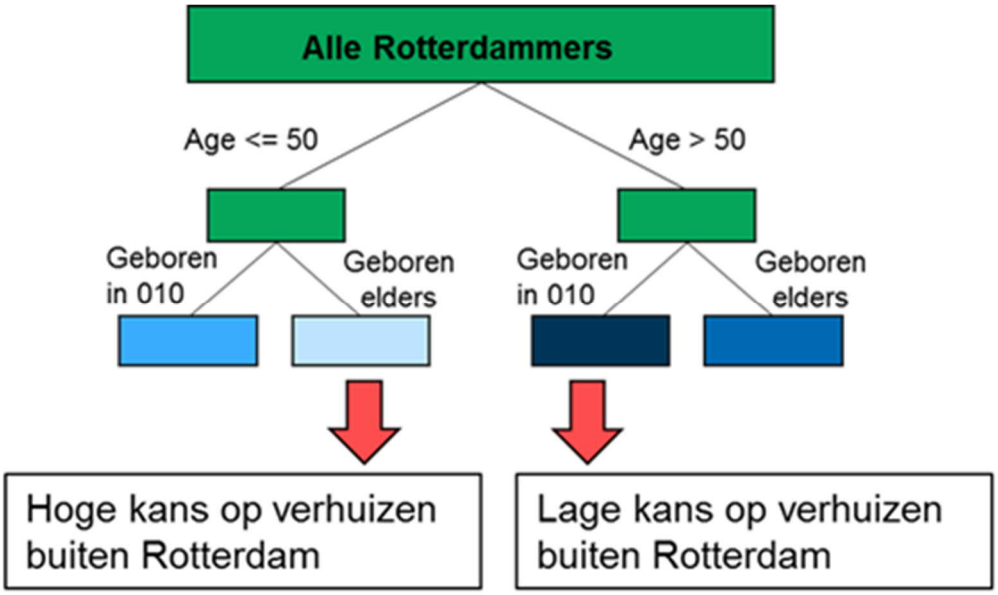
Hierna wordt volstaan met een aantal kernbegrippen en basisgedachten, die bij de toepassing van deze modellen bij de ontwikkeling van een algoritme een rol spelen.

Beslisboom[9]
Een beslisboom (of Decision Tree) is een modeltechniek in machine learning. Deze techniek begint met de totale groep en kiest vervolgens één variabele die het beste bijdraagt aan het maken van onderscheid binnen de groep. In ons geval, een variabele die het beste bijdraagt aan het maken van onderscheid tussen rechtmatigheid of onrechtmatigheid. In de onderliggende vertakking wordt weer een nieuwe variabele toegevoegd, waarmee weer nieuwe vertakkingen worden gemaakt.

Fictief voorbeeld:
Stel bijvoorbeeld dat we van alle Rotterdammers met behulp van een algoritme de kans willen voorspellen dat zij binnen de komende jaren buiten Rotterdam zullen verhuizen. Stel dat we in de gebruikte historische informatie terugzien dat de leeftijd hier een belangrijke voorspellende waarde heeft. Is de Rotterdammer al op hogere leeftijd, dan is de kans dat er nog een verhuizing buiten Rotterdam optreedt een stuk kleiner dan wanneer het een jongere Rotterdammer betreft.

Daarnaast blijkt de vraag of de inwoner van Rotterdam ook in Rotterdam is geboren, een variabele te zijn die extra bijdraagt aan het doen van een voorspelling. Vooral bij de jongere inwoner van Rotterdam die niet in Rotterdam geboren is, is de kans groter dat deze binnen enkele jaren weer Rotterdam verlaat. Juist bij de wat oudere inwoner die ook nog eens in Rotterdam 'geboren en getogen is' is de kans dat deze inwoner Rotterdam verlaat, aanmerkelijk kleiner.

In de figuur op de volgende pagina is dit voorbeeld uitgewerkt in een beslisboom.
De intensiteit van het blauw geeft hierin de hoogte van de kans weer
- hoe lichter de kleur blauw, hoe hoger de kans op verhuizing;
- naarmate het blauw intenser wordt, is de kans groter dat de inwoner niet buiten Rotterdam zal verhuizen.



Figuur 2: schematische weergave beslisboom 'kans op verhuizen buiten Rotterdam'.

[9] Zie Artificial Intelligence. A modern Approach. Third Edition van Stuart Russell en Peter Norvig, pagina 697.

Zo kunnen we (met hulp van berekeningen uitgevoerd door de computer) steeds op zoek naar extra variabelen die we 'onder aan de boom toevoegen' en die bijdragen aan een steeds betere voorspelling voor de kans op verhuizing buiten Rotterdam.

Ensemble[10]
Bij een ensemble maken we gebruik van verschillende modellen (hypotheses). Deze verschillende modellen kunnen dan op verschillende manieren met elkaar gecombineerd worden om tot gecombineerde voorspellingen te komen. De manier waarop de verschillende modellen worden gecombineerd, is afhankelijk van de gebruikte modelleertechniek (bijvoorbeeld Random forest of Gradient boosting, waarop we hierna wat dieper ingaan).

Een voorbeeld van een ensemble is dat meerdere beslisbomen (decision trees) worden gemaakt op hetzelfde onderwerp maar die er allemaal net iets anders uitzien.

Stel bijvoorbeeld (voor bovenstaand voorbeeld waarmee we de kans willen voorspellen van een verhuizing buiten Rotterdam) dat we een alternatieve beslisboom hebben gemaakt, waarbij we kijken naar de wijk waar de Rotterdammer nu woont en de gezinssamenstelling.

Vervolgens laten we beide beslisbomen een voorspelling doen van de kans op verhuizing. De voorspellingen uit beide beslisbomen worden gecombineerd tot één optimale voorspelling.

Gradient boosting[11]
Zoals eerder aangegeven, zijn er meerdere modelleertechnieken die kunnen worden gebruikt bij de ontwikkeling van een algoritme. De gebruikte modelleer-techniek kan van jaar tot jaar verschillen: we kiezen uiteindelijk voor *dat* model dat in enig jaar het beste slaagt in het doen van een goede voorspelling over (in ons geval) rechtmatige of onrechtmatige verstrekking van bijstand.
Omdat gradient boosting de techniek is die is gebruikt voor het in 2020 en 2021 gehanteerde model, wordt die techniek hier iets uitgebreider beschreven. Hopelijk geeft dit ook een basaal inzicht in hoe modelleertechnieken bij algoritmen werken; maar ook hoe complex dit vakgebied is.

Bij boosting maken we gebruik van meerdere beslisbomen; het is daarmee een ensemble methode. Daarbij zijn de verschillende beslisbomen bij boosting *een reactie op elkaar.* Dus de tweede beslisboom is gebaseerd op de performance van de eerste beslisboom. Performance is daarbij de mate waarin de beslisboom succesvol een voorspelling kan doen van (in ons geval) rechtmatigheid of onrechtmatigheid. Zo proberen we met elke nieuwe beslisboom de resultaten van de voorgaande verzameling beslisbomen iets te verbeteren. Uiteindelijk zijn er dus meerdere bomen gemaakt, waarvan de uitslag wordt gecombineerd tot één voorspelling voor de individuele uitkeringsgerechtigde.
Gradient boosting is specifieke vorm van boosting. Bij Gradient boosting wordt gestart met een beslisboom. De tweede beslisboom probeert vervolgens de fouten (onjuiste voorspellingen) van de eerste beslisboom te voorspellen. De verbeterde informatie van de tweede beslisboom wordt toegevoegd aan de voorspelkracht (performance) van de eerste beslisboom. Alleen wordt deze nieuwe informatie uit de tweede boom, slechts in een beperkte mate meegewogen, Dit getal, waarmee wordt bepaald in welke (beperkte) mate deze nieuwe informatie wordt meegenomen in het model heet de 'learning rate'.

Dit zorgt ervoor dat het model een klein stapje beter wordt bij gebruik van de tweede beslisboom. Op de nieuwe combinatie van beslisboom 1 en 2 wordt weer een nieuwe boom

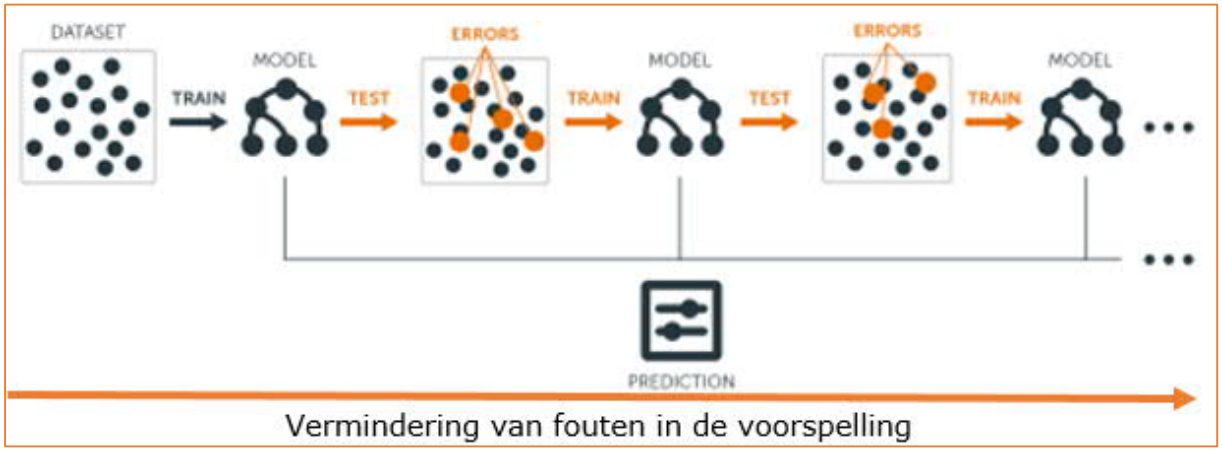
[10] Zie Artificial Intelligence. A modern Approach. Third Edition van Stuart Russell en Peter Norvig, pagina 748
[11] Voor een uitleg over boosting-technieken wordt verwezen naar Artificial Intelligence. A modern Approach. Third Edition van Stuart Russell en Peter Norvig, pagina 749

gemaakt om de fouten van deze combinatie te voorspellen. Ook deze derde beslisboom telt weer in een bepaalde verhouding (de learning rate) mee.

Dit gaat zo door totdat het model niet beter wordt door toevoeging van een nieuwe beslisboom. Dat is dus het geval wanneer een toevoeging van een nieuwe beslisboom, niet bijdraagt aan het doen van een betere voorspelling van (in ons geval) rechtmatigheid of onrechtmatigheid.
Het is echter ook mogelijk dat er vooraf een harde stop is aangegeven door de ontwikkelaar. Bijvoorbeeld voor het aantal beslisbomen dat kan worden gebruikt; of door een beperking in de tijd (er is niet meer tijd beschikbaar om verdergaande verbeteringen aan te brengen of de toevoegingen aan het model hebben zo'n beperkte invloed op de voorspellende kracht, dat het niet zinvol is om hier meer tijd of mensuren in te steken).

Uiteindelijk is er dan een volledig uitgewerkt model, waarin aan alle gebruikte features als gevolg van de hierboven beschreven stappen, een eigen relatieve belangrijkheid is toegekend. Deze relatieve belangrijkheid geeft aan in hoeverre de individuele feature die is meegenomen in het model, van belang is voor het doen van de voorspelling. Hoe hoger de relatieve belangrijkheid, des te meer weegt de feature mee bij het doen van de voorspelling. Een nadere uiteenzetting over de relatieve belangrijkheid is opgenomen in de bijlage.

In onderstaande figuur is een illustratie opgenomen van hoe de gradient boosting techniek werkt.



Figuur 3: schematische weergave van het proces van gradient boosting

Welke waarborgen zijn er dat het model goed wordt getraind?
Het ontwikkelen van het model gebeurt door ter zake deskundige en opgeleide medewerkers, die kennis hebben van en ervaring met de toepassing van de verschillende technieken. Bij het trainen van dergelijke modellen wordt gebruik maken van erkende programmatuur, opgesteld in ofwel de programmeertaal Python of R, hetgeen beiden internationaal erkende programmeertalen zijn voor de ontwikkeling van algoritmen. Het gaat hier om open source software, die door een internationale community van gebruikers wordt gebruikt en op die manier ook gevalideerd. De community van (deskundige) gebruikers zorgt er zelf voor dat fouten worden verwijderd en de software telkens wordt verbeterd.

Het ontwikkelen van het model gebeurt door ter zake deskundige en opgeleide medewerkers (data scientists), die kennis hebben van en ervaring met de toepassing van de verschillende technieken en gebruik maken van erkende programmatuur.

Hoe wordt het model getoetst?
We maken onderscheid tussen historische data, trainingsdata en testdata.

De historische data zijn alle gegevens die we hebben over uitkeringsgerechtigden waarbij

eerder een rechtmatigheidsonderzoek is uitgevoerd, waaronder de conclusie (rechtmatig of onrechtmatig).
Deze set splitsen we op in *trainingsdata* en *testdata*.

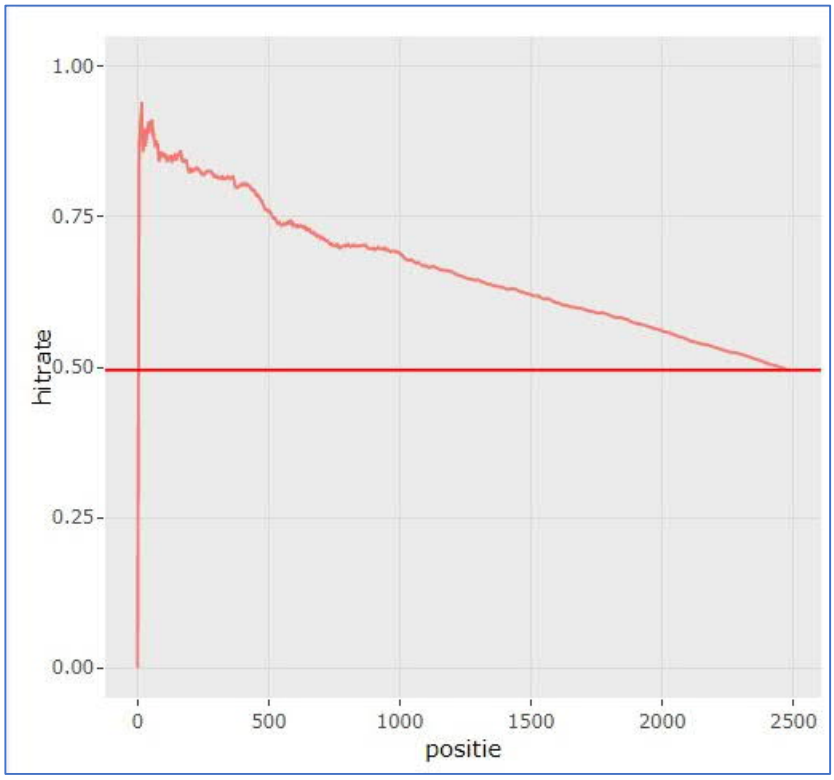
De trainingsdata bestaat uit ca. 70-80% van de historische data. Deze data gebruiken we om het algoritme te trainen om een model op te leveren dat in de hoogst mogelijke mate voorspellend is.

De testdata bestaat uit ca 20-30% van de historische data, die we *niet* betrekken bij de ontwikkeling van de trainingsdata. Van deze groep testdata weten we de uiteindelijke uitkomst (rechtmatig of onrechtmatig) vanuit het verleden. Nu gaan we kijken in welke mate het algoritme zoals dat is ontwikkeld op basis van de trainingsdata, ook in de testdata goed kan voorspellen of er sprake zal zijn van onrechtmatigheid of juist rechtmatigheid.

Om de performance van het model te kunnen toetsen, zijn twee aspecten van belang:
1) Accuracy = dit is het aandeel correct voorspelde uitkomsten. Het gaat dus in ons geval om het percentage van de gevallen waarin het algoritme correct kan voorspellen dat er (in de testdata) sprake is geweest van onrechtmatigheid of juist rechtmatigheid. Dat percentage ligt tussen 0 en 100%; een hoger percentage duidt op een model met een beter onderscheidingsvermogen (en is dus in staat om een betere voorspelling te doen).

2) Hitrate: hierbij worden de scores die het model binnen de testdata oplevert, geordend van hoog naar laag. Bovenaan staan dan de uitkeringsgerechtigden met het hoogste risico-inschattingsgetal (zie uitleg in hoofdstuk 4). Vervolgens wordt gekeken bij welk deel van de top-1, top-2, top-3 (etc.) op deze lijst, de uitkomst (onrechtmatigheid = ja) correct voorspeld is.

In onderstaande figuur is een voorbeeld van het verloop van de hitrate weergegeven:



Figuur 4: schematische weergave verloop hitrates

De horizontale as geeft de positite van de werkzoekende weer, geordend op basis van de hoogte van het risico-inschattingsgetal (van hoog naar laag). De hitrates worden dan als volgt berekend:

- We beginnen bij de top-1: was de werkelijke uitslag 'Onrechtmatig = Ja' dan is de hitrate in dat specifieke geval 100%, anders is het 0%. Dit is dan het 1e punt in de grafiek.
- Dan pakken we de 2e werkzoekende (met het op één-na-hoogste risico-getal) erbij: als de 1e en de 2e werkzoekenden uit de testdata in werkelijkheid allebei de uitslag 'Onrechtmatig = Ja' hadden, dan is de hitrate op het 2e punt ook 100%. Als ze allebei de uitslag ''Onrechtmatig = Nee' hadden, dan is de gezamenlijke hitrate 0%. Was de voorspelling in één van de twee gevallen goed, maar in het andere geval fout, dan is de totale hitrate 50%.
- Dan voegen we de 3e werkzoekende toe. Nu worden de mogelijke hitrates 100%, 66,67%, 33, 33%, of 0% (voor resp. 3, 2, 1 of 0 correcte voorspellingen).
- Etc.

In het begin van zo'n grafiek komen dus vaak grote sprongen voor (van 100 naar 50% bij voorbeeld), maar de fluctuatie wordt steeds kleiner. In principe is het een dalende lijn: naar mate het risico-inschattingsgetal daalt, zal immers ook minder vaak correct de voorspelling 'Onrechtmatig = Ja' worden gedaan.
Het is dan van belang dat het model vooral bij de gevallen waarin volgens het model sprake is van een hoge kans op onrechtmatigheid, correct heeft voorspeld of er sprake is van (on)rechtmatigheid. Deze uitkeringsgerechtigden hebben immers ook de grootste kans om te worden uitgenodigd voor een heronderzoek.

Met deze twee waarden (accuracy en hitrate) tezamen kan bepaald worden in hoeverre het model in staat is om correct te voorspellen of er sprake is van (on)rechtmatigheid.

De bruikbaarheid van het model wordt getoetst door de historische data te splitsen in trainingsdata en testdata. De effectiviteit van het model dat met de trainingsdata is ontwikkeld, wordt getoetst met de testdata. Hierbij vormen 'accuracy' en 'hitrate' belangrijke meetpunten.

<u>Samenvattend:</u>

- Het model wordt getraind om verbanden te vinden in de beschikbare data. Dat moet het model 'leren'.
- Daarvoor bestaan verschillende, wetenschappelijk onderbouwde en internationaal erkende technieken, die wisselend worden toegepast – er is géén sprake van één vast model dat jaarlijks opnieuw wordt gehanteerd.
- Het ontwikkelen van het model gebeurt door ter zake deskundige en opgeleide medewerkers, die kennis hebben van en ervaring met de toepassing van de verschillende technieken en gebruik maken van erkende programmatuur (software).
- De mate waarin het model slaagt om de juiste voorspelling te doen, wordt getoetst door het model dat op basis van de *trainingsdata* is ontwikkeld, te testen op *testdata*.

Hoofdstuk 4: De output

Welke output levert het algoritme op?
In hoofdstuk 3 is beschreven, hoe het model wordt 'getraind' om te komen tot optimale voorspellingen over de kans op rechtmatigheid of juist onrechtmatigheid per uitkeringsgerechtigde.

Het algoritme heeft nu dus 'geleerd' welke persoonskenmerken samenhangen met een hoge kans op onrechtmatigheid of juist rechtmatigheid. Dit heeft het algoritme geleerd op basis van historische data. Door nu te kijken in welke mate deze kenmerken zich voordoen bij de huidige groep van uitkeringsgerechtigden, kan er een risico-score (risico-inschattingsgetal) worden toegekend aan elke uitkeringsgerechtigde. Hierbij kijkt het model dus naar de mate waarin de 'waarden' van de individuele uitkeringsgerechtigde matchen met de waarden van de variabelen uit de historische data.

Deze score varieert van 1 (grotere kans op onrechtmatigheid) tot 0 (kleinere kans op onrechtmatigheid).

Het risico-inschattingsmodel zorgt voor een lijst waarop de huidige uitkeringsgerechtigden worden weergegeven in een rangorde van 1 (grotere kans op onrechtmatigheid op basis van het algoritme) tot 0 (kleinere kans op onrechtmatigheid).

Hoe vindt de selectie voor een heronderzoek plaats?
Vervolgens worden die uitkeringsgerechtigden geselecteerd voor een heronderzoek op basis van het risico-inschattingsmodel, waarvoor de score het dichtst bij 1 ligt. Dit zijn dus de uitkeringsgerechtigden waarbij het model een relatief hoge kans op onrechtmatigheid heeft ingeschat.

Vooraf wordt een besluit genomen over de wijze waarop het in enig jaar te verrichten aantal heronderzoeken wordt verdeeld. Zo was voor het jaar 2021 gepland dat er 7.500 heronderzoeken zouden worden uitgevoerd[12], waarvan 2.000 geselecteerd op basis van het risico-inschattingsmodel. Dit betekent dus in principe dat de eerste 2.000 uitkeringsgerechtigden op de lijst worden geselecteerd, waarbij de lijst is geordend op basis van het risico-inschattingsgetal (dus van 1 tot 0).

Overigens vindt op deze uitkomst ook weer een verdere filtering plaats, bijvoorbeeld om ervoor te zorgen dat een individuele uitkeringsgerechtigde niet twee keer in een kort tijdsbestek kan worden uitgenodigd voor een heronderzoeksgesprek.
Voor deze filtering wordt verwezen naar de (op dit moment nog op te leveren) procesbeschrijving. Dit valt dus buiten het bestek van deze notitie.

Het aantal heronderzoeken dat op basis van het risico-model wordt geselecteerd, wordt jaarlijks opnieuw vastgesteld. Dit valt buiten het bestek van deze notitie

.

[12] Inmiddels is duidelijk dat deze planning, onder meer door de voortdurende Covid19-crisis, niet zal worden behaald.

BIJLAGE: Uitleg variabelen en features

Revelante begrippen:

Variabele
Variabele zoals deze uit de brondata zijn verkregen voordat ze naar features zijn omgezet.

Feature
Verwerking van een variabele tot een weergave waarmee deze het model ingaat. Dit kan bijvoorbeeld door dummyficatie; dit betekent dat je voor elke categorie die een variabele kan hebben een nieuwe variabele aanmaakt die 1 (waar) of 0 (onwaar) is. In het onderstaande voorbeeld woont persoon 1 in Feijenoord, dus in de featureweergave heeft deze een 1 variabele “adres_recentste_wijk_feijenoord”, en een 0 in de andere kolommen.

Variabele weergave	
Persoon	Adres_recentste_wijk
1	Feijenoord
2	IJsselmonde
3	Delfshaven

Feature weergave			
Persoon	Adres_ recentste_ wijk_ feijenoord	Adres_ recentste_ wijk_ ijsselmonde	Adres_ recentste_ wijk_ delfshaven
1	1	0	0
2	0	1	0
3	0	0	1

Andere variabelen kunnen tekst bevatten. Dergelijke tekstuele variabelen verwerken we tot features door alleen te kijken naar of de variabele tekst bevat. De betreffende tekst wordt niet in tekstuele vorm in het model gebruikt. In het onderstaande voorbeeld hebben persoon 1 en persoon 3 tekstuele informatie in de kolom uiterlijke_verzorging_opm, deze krijgen daarom in de feature weergave een 1 in deze kolom. Omdat persoon 2 geen informatie in deze kolom heeft staan heeft deze persoon waarde 0.

Variabele weergave		
Persoon	Uiterlijke_ verzorging_ opm	Motivatie_opm
1	"Verzorgd"	"Client neemt weinig initiatieven"
2		"Solliciteert actief"
3	"Soms onverzorgd"	

Feature weergave		
Persoon	Uiterlijke_ verzorging_ opm	Motivatie_ opm
1	1	1
2	0	1
3	1	0

Relatieve belangrijkheid:

De relatieve belangrijkheid geeft aan hoe belangrijk een feature is in het model. De belangrijkheid van features zijn altijd relatief ten opzichte van elkaar. De belangrijkste feature heeft de index 100 en de andere features een kleiner getal volgens hun verhouding. Verdere uitleg over de berekening van de relatieve belangrijkheid is opgenomen aan het einde van deze bijlage.

Termen die in featurenaam voor kunnen komen:

Mean

Het gemiddelde

Boolean

Of een variabele waar of onwaar is (1,0)

Other

Bij het maken van categorieën, worden niet veel voorkomende categorieën samen in de categorie other (overig) samengevoegd. Niet veel voorkomende categorieën zijn categorieën die bij minder dan 2% van de gevallen voorkomen. In het onderstaande voorbeeld komen wijken Centrum en Kralingen-Crooswijk maar één keer voor. Deze worden dan samengevoegd in de feature adres_recentste_wijk_other, waar personen 8 en 12 waarde 1 hebben.

Variabele weergave	
Persoon	adres_ recentste_ wijk
1	Feijenoord
2	Delfshaven
3	Delfshaven
4	Delfshaven
5	Feijenoord
6	Feijenoord
7	Delfshaven
8	Kralingen-Crooswijk
9	Feijenoord
10	Feijenoord
11	Delfshaven
12	Centrum

Feature weergave			
Persoon	adres_ recentste_ wijk_ feijenoord	adres_ recentste_ wijk_ delfshaven	adres_ recentste_ wijk_ other
1	1	0	0
2	0	1	0
3	0	1	0
4	0	1	0
5	1	0	0
6	1	0	0
7	0	1	0
8	0	0	1
9	1	0	0
10	1	0	0
11	0	1	0
12	0	0	1

Berekenen relatieve belangrijkheid

Relatieve belangrijkheid geeft zoals eerder uitgelegd aan hoe belangrijk elke feature is voor het doen van de voorspelling - in dit geval een verwachte kans op (on)rechtmatigheid bij het verstrekken van uitkering. Hoe dichter de relatieve belangrijkheid bij 1 ligt, des te belangrijker is de betreffende feature voor het doen van de voorspelling.

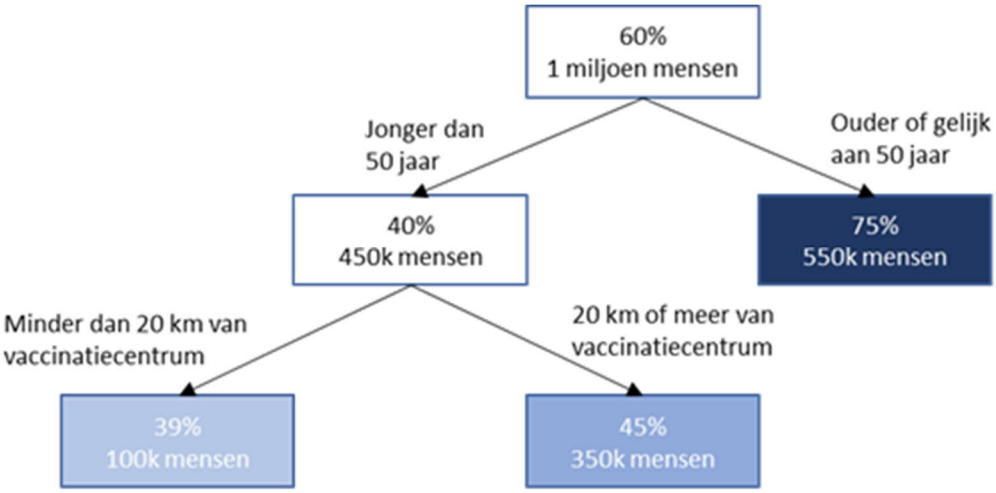
Relatieve belangrijkheid binnen een beslisboom wordt berekend op basis van twee factoren:

1. In hoeverre kan er op basis van de feature correct onderscheid worden gemaakt met betrekking tot de voorspelling die we willen doen (in dit geval tussen uitkeringsgerechtigden waarbij sprake is van onrechtmatigheid dan wel rechtmatigheid)?
2. Om hoeveel voorspellingen (in dit geval: aantal uitkeringsgerechtigden) gaat dit?

Dit wordt bepaald op basis van gegevens uit de beschikbare trainingsdata.

Fictief voorbeeld:
We willen van alle Rotterdamse inwoners voorspellen wie zich zal gaan laten vaccineren tegen Corona. Uit landelijke ervaringscijfers is bekend dat gemiddeld 60% van alle uitgenodigde personen zich laat vaccineren. Bij het bouwen van het model zal worden gezocht naar kenmerken op basis waarvan we zo goed mogelijk kunnen voorspellen wie tot welke groep (wel vaccineren / niet vaccineren) zal behoren. We zoeken dus naar features (kenmerken) met een hoge onderscheidende waarde.

We kennen van 1 miljoen landelijk gevaccineerden een aantal kenmerken en komen op basis daarvan tot de volgende beslisboom:



We zien dat de variabele *leeftijd* een grote voorspellende waarde heeft. De feature 'leeftijd 50 jaar of meer' krijgt in ons model een hoge relatieve belangrijkheid, omdat:

1. Leeftijd blijkbaar een grote invloed heeft op de kans op vaccinatie, immers:
   - In de groep jonger dan 50 jaar vaccineert zich 40%
   - In de groep ouder of gelijk aan 50 jaar vaccineert zich 75%

   Dat is een vrij groot verschil; deze feature heeft dus een *grote* onderscheidende waarde.
2. Deze splitsing konden we toepassen op de volledige groep van 1 miljoen mensen. Daarom zal deze feature ook zwaarder meetellen in ons model.

We zien dat de variabele *afstand* tussen woonplaats en priklocatie vervolgens helpt om ons model verder te verfijnen. De feature 'Afstand <= 20 kilometer' krijgt echter (in vergelijking met de feature 'leeftijd 50 jaar of meer') een lagere relatieve belangrijkheid in het model omdat:

1. We met de splitsing op basis van een afstand < of > dan 20 kilometer slechts een relatief *klein* onderscheiding weten aan te brengen:
   - In de groep die verder dan 20 kilometer van de priklocatie woont vaccineert zich 45%
   - In de groep die binnen een straal van 20 kilometer van de priklocatie woont vaccineert zich 39%.

   Het verschil tussen 45% en 39% is vrij klein, vergeleken met de onderscheidende waarde van de feature 'leeftijd 50 jaar of meer'. Deze feature heeft dus een *kleinere* onderscheidende waarde en krijgt daarom een lagere relatieve belangrijkheid toegekend in het model.
2. Dit onderscheid hebben we hier alleen losgelaten op de groep jonger dan 50 jaar. Deze groep (450k) is relatief klein vergeleken met de groep waarop we de feature

'leeftijd 50 jaar of meer' konden toepassen. Ook daarom is de relatieve belangrijkheid van deze feature lager.

Voor de techniek van gradient boosting model zijn er nog twee belangrijke toevoegingen:

- Zoals in hoofdstuk 3 bij de uitleg van gradient boosting is toegelicht, is er sprake van een 'learning rate'. Het model bestaat uit meerdere bomen, waarbij het effect van volgende (lagere) beslisbomen in beperkt(ere) mate wordt meegewogen[13].
- Een feature kan op meerdere plekken terugkomen in de boom. In bovenstaand voorbeeld kan de groep van 50 jaar of ouder wellicht ook verder worden onderverdeeld op basis van afstand tot de priklocatie. Dat heeft alleen meerwaarde als we zien dat de afstand tot de priklocatie ook bij 50+-ers van invloed is op de vaccinatiebereidheid. Heeft dat inderdaad ook bij 50+-ers een effect op de kans op vaccineren, dan kan ook hier worden bepaald welke percentages (vaccineren / niet vaccineren) daarbij horen. De effecten binnen de groep jonger dan 50 jaar en binnen de groep 50+ worden dan bij elkaar opgeteld om tot één totaalcijfer te komen waarmee de relatieve belangrijkheid van de feature 'Afstand <= 20 kilometer' wordt uitgedrukt.

[13] Zie de uitleg over het begrip 'learning rate' in hoofdstuk 3.