Risicoprofilering via nieuwe technologieën: juridische beoordeling vraagt meer dan het kennen van het recht

drs. A.T.H. van der Linden[1]

1. Inleiding

In de afgelopen decennia zijn grote stappen gezet binnen het domein kunstmatige intelligentie, inclusief het subdomein machine learning. De ontwikkelingen gaan niet voorbij aan Nederlandse overheidsinstanties. Zij zetten nieuwe technologieën, zoals machine learning, steeds vaker in als risicoprofilering in de handhaving. Nieuwe technologieën verdiepen de mensenrechtelijke uitdagingen ten opzichte van rule-based algoritmische systemen, bijvoorbeeld omdat het bij nieuwe technologieën lastiger is om discriminatie op te sporen.[2] Er zijn tal van voorbeelden waarin de technologische inzet (waarschijnlijk) niet binnen het wettelijke kader bleef. Denk bijvoorbeeld aan de gebruikte technologieën bij Binnenlandse Zaken (gebruikt om visumaanvragen te beoordelen)[3], DUO (gebruikt bij de fraudebestrijding met studiefinanciering)[4], diverse gemeenten (gebruikt om bijstandsfraude op te sporen)[5] en bij de Belastingdienst (risicomodel toeslagen met discriminerende variabele "Nederlander ja/nee")[6]. In deze position paper ga ik vanwege de omvang alleen in op de werking alsmede de juridische toetsing van supervised[7] machine learning modellen.

2. Discriminatie

Discrimineren betekent het maken van een verboden onderscheid. Het gemaakte onderscheid is verboden als er geen objectieve rechtvaardiging voor bestaat. Een onderscheid is wel objectief gerechtvaardigd als er sprake is van een legitiem doel, het onderscheid proportioneel is, en het onderscheid niet op een minder ingrijpende wijze kan worden gemaakt om hetzelfde doel te bereiken. Het bestaan van een statistisch verband is geen toereikende reden om daarop onderscheid te maken in de handhaving. Er is sprake van directe discriminatie als er uitsluitend onderscheid wordt gemaakt op basis van een beschermd kenmerk zoals geslacht, godsdienst, nationaliteit of ras. In de handhaving mag alleen op basis van een beschermd kenmerk onderscheid worden gemaakt als dat, vanuit de wettelijke regeling bezien, een relevante factor is. Er is sprake van indirecte discriminatie wanneer een ogenschijnlijk neutrale handelwijze personen met een bepaald kenmerk in het bijzonder treft en dat onderscheid niet objectief gerechtvaardigd is.

Het wettelijke kader[8] is wat mij betreft toereikend.[9] Met deze position paper wil ik laten zien dat de volledige kennis van een duidelijk wettelijk kader nog niet hoeft te leiden tot een juiste rechtsbeoordeling. Fiscalisten en juristen zijn traditioneel "*niet opgeleid in dataherkenning, dataselectie en statistiek*".[10] Diezelfde kennis, in combinatie met de kennis van de werking van de ingezette technologie, is mijns inziens wel vereist voor de rechtsbeoordeling.[11]

Ondanks de wettelijke waarborgen die er zijn om inzicht te krijgen in de selectiemethodieken[12], is het voor burgers moeilijk in te schatten of ze worden gediscrimineerd. Dat vind ik een kritiek punt in het wettelijke kader c.q. uitvoeringskader. Pas toen toeslagenouders met elkaar in contact kwamen,

---

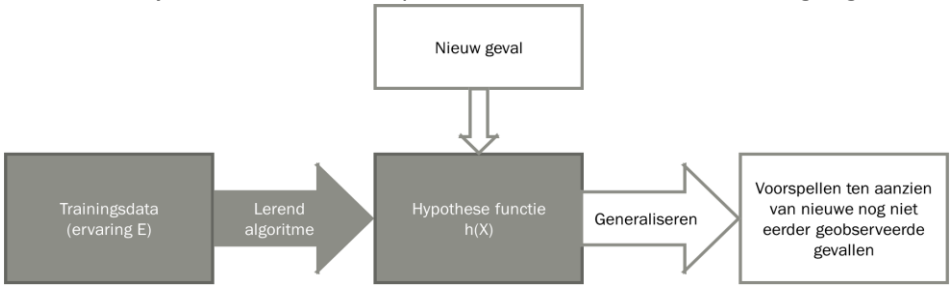
[1] De auteur is als docent/coördinator Tax & Technology verbonden aan Tilburg University. Vanuit die hoedanigheid volgde de uitnodiging voor het rondetafelgesprek. Daarnaast is hij als specialistisch adviseur Tax & Technology verbonden aan de Belastingdienst. De bijdrage aan het rondetafelgesprek, inclusief deze zogenoemde position paper, is op persoonlijke titel.
[2] Rathenau Instituut, Algoritmes Afwegen, p. 66-67.
[3] https://www.nrc.nl/nieuws/2023/05/01/minister-moet-uitleg-geven-over-algoritme-voor-visa-a4163510.
[4] https://nos.nl/artikel/2481624-autoriteit-persoonsgegevens-onderzoekt-fraudesysteem-duo.
[5] https://www.autoriteitpersoonsgegevens.nl/actueel/ap-wil-opheldering-over-fraude-algoritme-gemeenten.
[6] De Autoriteit Persoonsgegevens toetste of de gegevensverwerking behoorlijk was in het licht van artikel 5, lid 1, aanhef en sub a, AVG. Het gehanteerde toetsingskader is gelijk van toepassing op artikel 26 IVBPR, artikel 1 van Protocol nr. 12 bij het EVRM en artikel 1 van de Grondwet. Zie Autoriteit Persoonsgegevens, Belastingdienst/Toeslagen De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag, onderdelen 3.7 t/m 3.7.3.
[7] Supervised machine learning is op dit moment de meest succesvolle vorm van machine learning. Zie Y. LeCunn, Y. Bengio & G. Hinton, Deep Learning, Nature, Vol. 521 (2015), p. 436. Vergelijk C. Müller, S. Guido, Introduction to Machine Learning with Python, O'Reilly: 2017, p. 25.
[8] Het grondwettelijke discriminatieverbod (artikel 1 Grondwet) is nader uitgewerkt in verschillende andere wetten, zoals bijvoorbeeld de Algemene Wet Gelijke Behandeling.
[9] Er zijn ook diverse hulpmiddelen die kunnen helpen bij het toetsen op non-discriminatie. Zie bijvoorbeeld Impact Assessment Mensenrechten en Algoritmes, de handreiking non-discriminatie by design en discriminatie door risicoprofielen – een mensenrechtelijk toetsingskader van het College voor de Rechten van de Mens.
[10] A.F.M.Q. Beukers-Van Dooren, Waarom je als rechter het WFR moet lezen, WFR 2022/17, paragraaf 4.
[11] Vergelijk mijn eerdere betoog in A.T.H. van der Linden, Het WFR is al 150 jaar de "brandstof" van de fiscalist, WFR 2022/119, paragraaf 2.
[12] Vergelijk HR 17 augustus 2018, ECLI:NL:HR:2018:1316. Vergelijk ook HR 10 december 2021, ECLI:NL:HR:2021:1748.

bleek dat mensen met een migratieachtergrond oververtegenwoordigd waren. Ook voor juridische dienstverleners geldt dat het recht kennen niet altijd voldoende is voor een juiste rechtstoepassing.

3. De werking van supervised machine learning[13]

Machine learning is een techniek waarbij een model wordt getraind om een bepaalde taak T uit te voeren. In plaats van, zoals bij traditionele software, alle regels expliciet te programmeren, worden deze regels bij machine learning afgeleid uit een dataset met behulp van wiskundige en statistische methoden. Overheidsinstanties kunnen supervised machine learning modellen bijvoorbeeld inzetten om potentieel foutieve aanvragen of aangiften te selecteren. Het misschien bekendste voorbeeld van machine learning bij een overheidsinstantie zijn de risicomodellen bij de Belastingdienst.[14]

De kern van supervised machine learning ga ik uitleggen aan de hand van het volgende versimpelde processchema. Daarbij richt ik me vooral op risicomodellen, maar de werking is generiek.



3.1 Trainingsdata

De trainingsdata zijn de grondstof van een machine learning model. De trainingsdata bevatten de observaties waarvan het model moet leren. Volgens het rapport van de Algemene Rekenkamer worden risicomodellen van de Belastingdienst getraind aan de hand van aangiften waarvan de controle-uitkomst met een grote mate van zekerheid bekend is.[15] Het lijkt dan voornamelijk om aangiften te gaan die zijn behandeld als gevolg van een steekproef.[16]

De trainingsdata zijn op te delen in twee delen: de onafhankelijke variabelen (‘features’) en de afhankelijke variabele (‘labels’). De onafhankelijke variabelen zijn potentieel alle data of kenmerken van de aangifte en de belastingplichtige die de aangifte heeft ingediend.[17] De afhankelijke variabele is de controle-uitkomst van de aangifte als doelvariabele: ‘juist’ of ‘onjuist’.[18]

3.2 Het lerende algoritme

Het lerende algoritme legt de patronen en de verbanden tussen de onafhankelijke en afhankelijke variabelen. Het machine learning model kan op geautomatiseerde wijze informatie extraheren door in de trainingsdata patronen te herkennen. Expliciete instructies van een mens zijn dan niet meer nodig om taak T te bereiken.

Het is belangrijk om te benoemen dat het lerende algoritme sterk gebruik maakt van statistische kansrekeningen.[19] Bij het risicomodel toeslagen volgde de weging van de variabele “Nederlander ja/nee” dus uit statistische kansberekeningen. Machine learning belooft patronen en gevolgtrekkingen te vinden die waarschijnlijk correct zijn voor de meeste observaties, maar deze patronen en gevolgtrekkingen zijn vaak slechts toevallig gerelateerd aan mensachtige logica.[20]

---

[13] Het betreft hier een bespreking op hoofdlijnen. A.T.H. van der Linden & R. Hein, The Impact of Technology (in: Tax Assurance), Deventer: Wolters Kluwer 2022, p. 147-157.
[14] Algemene Rekenkamer, Datagedreven selectie van aangiften door de Belastingdienst, p. 28-29. De term machine learning komt niet met zoveel woorden in het rapport van de Algemene Rekenkamer terug, maar de methodiek en werking zijn volledig in overeenstemming met de werking van machine learning.
[15] Het rapport van de Algemene Rekenkamer ziet op de inkomstenbelasting en de omzetbelasting. Daarom spreek ik van een aangifte en niet van een aanvraag zoals bij een toeslag. Zie Algemene Rekenkamer, Datagedreven selectie van aangiften door de Belastingdienst, p. 5.
[16] Algemene Rekenkamer, Datagedreven selectie van aangiften door de Belastingdienst, p. 28.
[17] Algemene Rekenkamer, Datagedreven selectie van aangiften door de Belastingdienst, p. 5.
[18] Algemene Rekenkamer, Datagedreven selectie van aangiften door de Belastingdienst, p. 28-29.
[19] I. Goodfellow, Y. Bengio & A. Courville, Deep Learning, MIT Press: 2017, p. 52.
[20] I. Goodfellow, Y. Bengio & A. Courville, Deep Learning, MIT Press: 2017, p. 113.

Juridisch kan dat uitermate problematisch zijn, omdat dan (ook) andere dingen worden gewogen dan de fiscaal relevante criteria.

3.3 De hypothese functie *h(x)*

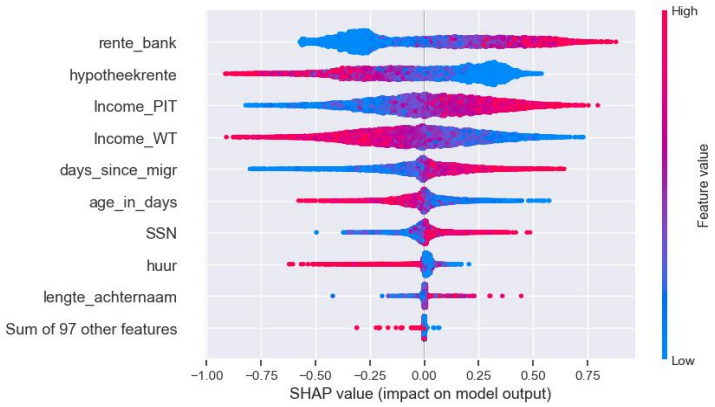
Het lerende algoritme heeft op basis van de relatie tussen de onafhankelijke variabelen aan de ene kant en de afhankelijke variabele aan de andere kant geleerd wat de kenmerken zijn van een juiste en een onjuiste aangifte. De blauwdruk wordt dus gemaakt op basis van data uit het verleden. De statistische aannames die achter het proces schuilen, zijn in het fiscale recht minder reëel dan bij alledaagse toepassingen van machine learning, zoals FaceID, spamfilters, vertalingen en objectherkenning.[21]

3.4 Het nieuwe geval: generaliseren en voorspellen

Het nieuwe geval is een nieuwe aangifte. De binnengekomen aangifte wordt tegen de hypothese functie *h(x)* aangehouden. Op basis van de onafhankelijke variabelen van het nieuwe geval, dat zijn volgens het rapport van de Algemene Rekenkamer alle data en kenmerken van de aangifte en de belastingplichtige, volgt een voorspelling. De voorspelling is de procentuele kans dat de aangifte een fout bevat.[22] De kans op een fout in de aangifte wordt groter naarmate de onafhankelijke variabelen van de ingediende aangifte meer overeenkomsten vertonen met de data en kenmerken van gecorrigeerde aangiften.

3.5 Het gemaakte onderscheid: welke variabelen gebruikt het model?

Een veelgebruikte methode om te achterhalen welke onafhankelijke variabelen een model precies gebruikt, en hoe die onafhankelijke variabelen aan de risicoscore bijdragen, is via Shapley-waarden.[23] Op basis van een synthetische dataset is in de volgende figuur (rechts) een voorbeeld opgenomen hoe Shapley-waarden eruitzien.[24]



Het is niet eenvoudig om de werking van Shapley-waarden te duiden als je ze voor het eerst ziet. Toch doe ik een poging, omdat het relevant is voor de juridische duiding van de variabelen. In dit voorbeeld staat een “1” voor een waarschijnlijk juiste aangifte en een “0” voor een waarschijnlijk onjuiste aangifte. Op de verticale as zie je welke variabelen het model gebruikt. Op de horizontale as zie je hoe de variabelen bijdragen aan de risicoscore. Een onafhankelijke variabele heeft een kleur die varieert van blauw tot donkerrood. Hoe blauwer de kleur bij een variabele is, hoe relatief lager de waarde van de variabele over het gehele spectrum bezien is. Andersom geldt hetzelfde, hoe roder de kleur bij een variabele is, hoe relatief hoger de waarde van de variabele over het gehele spectrum bezien is. Als iemand bijvoorbeeld een relatief korte achternaam heeft, dan is de kleur in dit voorbeeld blauwer en andersom. Een korte achternaam verlaagt in dit voorbeeld de score, waardoor de kans op een onjuiste aangifte volgens het model toeneemt.

Het machine learning model lijkt in eerste instantie geen onderscheid te maken op bijvoorbeeld iemands afkomst, omdat dit niet expliciet is meegegeven als inputgegeven. Toch is het tegendeel

---

[21] Het is belangrijk om te vermelden dat bij machine learning wordt aangenomen dat de data onafhankelijk en identiek is gedistribueerd (i.i.d. assumption). Deze aanname is vanuit fiscaal oogpunt problematisch, omdat de kans op een onjuiste aangifte voortdurend verandert door bijvoorbeeld wijzigingen in wetgeving, nieuwe rechterlijke uitspraken en de invulling van open normen. Hierdoor verandert de relatie tussen de onafhankelijke en afhankelijke variabelen continu. In tegenstelling tot grote technologiebedrijven hebben veel overheidsorganisaties geen continue datastroom, waardoor modellen vaak pas veel later kunnen worden aangepast aan de gewijzigde kansverdelingen. Zie uitgebreider A.T.H. van der Linden & R. Hein, The Impact of Technology (in: Tax Assurance), Deventer: Wolters Kluwer 2022, p. 152-153.

[22] Algemene Rekenkamer, Datagedreven selectie van aangiften door de Belastingdienst, p. 8.

[23] De stap naar Shapley-waarden is meestal een vervolgstap, want het bekijken van de coëfficiënten geeft al een eerste indicatie. Hier ga ik in het kader van de omvang niet nader op in.

[24] In verband met tijdsgebrek komen zowel Engelse als Nederlandse termen in de figuur terug.

waar. Dit synthetische model maakt op diverse manieren onderscheid op basis van iemands afkomst, via zogenoemde *proxies*.[25] Ik licht er enkele voorbeelden uit:

- De lengte van de achternaam kan iets zeggen over de herkomst van een persoon. Typisch Nederlandse achternamen zijn gemiddeld bezien langer, bijvoorbeeld als gevolg van tussenvoegsels. In dit voorbeeld draagt een langere achternaam bij aan de voorspelling dat de aangifte juist is en andersom.
- Hetzelfde geldt voor huur. Mensen met een migratieachtergrond huren statistisch bezien vaker een woning dan mensen zonder migratieachtergrond.[26] In dit voorbeeld draagt het huren van een woning (rood) negatief bij aan de score, waardoor de selectiekans voor huurders in dit voorbeeld hoger is en andersom.
- Ook een samenstel van variabelen kan leiden tot het maken van onderscheid op basis van iemands afkomst. In dit voorbeeld staat de variabele dagen sinds migratie (days_since_migr) voor het aantal dagen sinds iemand is aangemeld bij de gemeente. Als iemand vlak na de geboorte is aangemeld bij de gemeente, dan is de hoogte van de variabele ongeveer gelijk aan de leeftijd in dagen van deze persoon (age_in_days). Voor de risicoscore balanceren zij elkaar dan uit. Een hoge waarde voor leeftijd in dagen, maar een lage waarde voor dagen sinds migratie zorgt per saldo voor een negatieve bijdrage aan de score. Dit samenstel leidt tot een grotere selectiekans voor mensen met een discrepantie tussen de leeftijd in dagen en het moment van aanmelding bij de gemeente, zoals bij mensen met een migratieachtergrond.

Het klinkt paradoxaal, maar je hebt het verboden kenmerk nodig om te controleren of het model op indirecte wijze onderscheid maakt naar het betreffende verboden kenmerk. Je traint dan een apart model met als afhankelijke variabele het verboden kenmerk. Op die manier kun je controleren of er variabelen bestaan die (sterk) correleren met een verboden kenmerk.

4. Conclusie

Het juridische kader is voldoende helder. Een juiste rechtstoepassing vraagt meer dan het kennen van het recht. Zelfs als je precies weet hoe het discriminatieverbod juridisch in elkaar steekt, kan een gebrek aan kennis wat betreft de technologische (ver-)werking tot een onjuiste juridische beoordeling leiden. Dat juristen niet altijd goed begrijpen wat er onder de motorkap van de technologische verwerking gebeurt, omdat ze daar vaak niet in geschoold zijn, is een mogelijke oorzaak waarom het bij diverse overheidsinstanties misgaat qua technologische inzet. Er is een combinatie van technische en domeinkennis nodig om tot waardevolle toepassingen te komen, aldus de Wetenschappelijke Raad voor het Regeringsbeleid.[27]

Daarnaast blijkt uit deze position paper dat machine learning modellen ook (indirect) onderscheid kunnen maken op verboden kenmerken, terwijl dat niet altijd direct duidelijk is. Het is bij machine learning als selectiemethode mijns inziens problematisch dat er op basis van het verleden een soort profiel van een onjuiste aangifte wordt opgesteld.[28] Als jouw kenmerken bij het indienen van de aangifte, in combinatie met de kenmerken van de aangifte, in hoge mate overeenkomen met het profiel van een onjuiste aangifte, dan is de kans op behandeling groter. Identieke fouten leiden dan niet tot dezelfde selectiekans, omdat meer wordt meegewogen dan alleen de feitelijke gedragingen. Puur op basis van statistische overeenkomsten hadden mensen waarvoor gold “Nederlander nee” een grotere selectiekans dan mensen waarvoor gold “Nederlander ja”.[29]

---

[25] Vergelijk Algemene Rekenkamer, Datagedreven selectie van aangiften door de Belastingdienst, p. 19.
[26] Vergelijk https://www.cbs.nl/nl-nl/achtergrond/2013/15/ruim-600-duizend-corporatiewoningen-bewoond-door-allochtonen#:~:text=Bijn%2040%20procent%20van%20de,corporatie%20of%20een%20gemeentelijk%20woningbedrijf.
[27] Wetenschappelijke Raad voor het Regeringsbeleid (2016), Big data in een vrije en veilige samenleving, rapport nummer 95, p. 86.
[28] Zie tevens voetnoot 21 waarom de aannames in dit juridische proces vaak onrealistisch zijn.
[29] Binnen het risicomodel toeslagen. Het bestaan van een statistisch verband is geen toereikende reden om onderscheid te maken in de controlestrategie.