Selectie met AI: meer of minder eerlijk dan traditionele selectie methoden?

December 2020

Jacqueline van Breemen NOA BV

In opdracht van NSvP

Met medewerking van 3DUniversum

Het gebruik van Artificiële Intelligentie (AI) bij de selectie van studenten en werknemers is nog niet in volle gang, maar er is veel belangstelling om deze techniek toe te passen. Er is echter nog weinig onderzoek gedaan naar de consequenties van selectie met behulp van AI op de diversiteit van de geselecteerden. Er zijn aanwijzingen dat bestaande bias in de samenstelling van het huidige studenten of werknemersbestand overgenomen of zelfs versterkt wordt door het gebruik van AI bij de selectie van nieuwe studenten en werknemers.

Ook is er de meer fundamentelere vraag of selectie met modellen gebaseerd op psychometrische indicatoren werk of studiesucces beter of slechter voorspellen dan een AI model gebaseerd op meer algemenere gegevens. Voor de huidige recruitmentpraktijk, en voor assessoren en psychologen werkzaam in dit veld, heeft het antwoord op deze vraag mogelijk grote consequenties. AI modellen zijn efficiënter, en waarschijnlijk ook kosten-effectiever. De uitvoering van deze modellen wordt gedaan door data-scientists, niet door psychologen. Ook zijn data-scientists niet gehouden aan een beroepscode voor wat betreft integriteit, discriminatie, e.d.

Ten derde wordt er vanuit instanties die zich bezighouden met wettelijke en maatschappelijke consequenties (zoals bijv. Data&Society en de Europese Commissie) opgeroepen om eerlijke en transparante AI methoden te gebruiken. Echter, om deze te ontwikkelen is kennis nodig over wanneer, en op welke manier, een AI (selectie)methode verschilt van de meer traditionele methodes.

De vraag is dus hoe groot deze bias van AI voor recruitment en selectie is, en welke groepen aankomend werknemers en studenten nu precies voor- of nadeel ondervinden? Voordat er eerlijke en transparante AI methoden kunnen worden ontwikkeld moet er eerst onderzocht worden hoe meer of minder eerlijk AI is, vergeleken met de meer traditionele selectie methoden. Daarom hebben we een vergelijkend onderzoek uitgevoerd, waarbij we de bias vergelijken tussen de meer traditionele selectie methoden en AI selectie.

1. Algoritmes

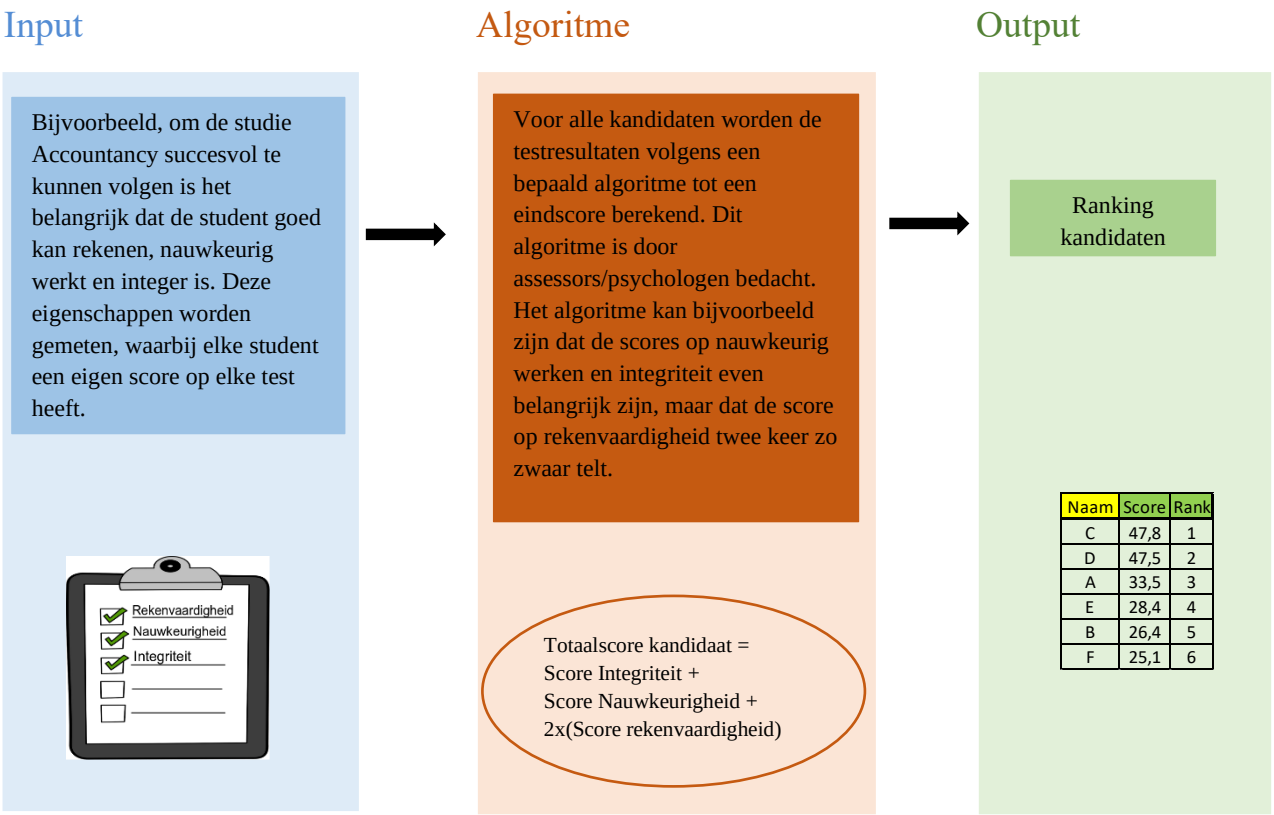
Bij selectie van studenten voor een studie, en sollicitanten voor een baan, worden vaak algoritmes ingezet. Een algoritme is niets anders dan een set regels die gevolgd worden om een probleem op te lossen. Met behulp van het algoritme probeer je een voorspelling te doen over welke selectiekandidaten succesvol kunnen zijn. Bij de selectie van studenten en sollicitanten zijn eigenschappen zoals bijvoorbeeld cognitieve vaardigheden en persoonlijkheidskenmerken belangrijk voor succes. Deze eigenschappen worden gemeten met, als het goed is, gevalideerde psychometrische testinstrumenten. De resultaten van deze tests worden doormiddel van een algoritme gecombineerd in een eindscore. Vooraf is bepaald aan welke eindscore een kandidaat minimaal moet voldoen, en dus of de student of sollicitant geschikt is.

Op dit moment worden deze selectiealgoritmes ontworpen door eerst een theoretisch model te maken. In dit model worden de cognitieve vaardigheden en persoonlijkheidskenmerken die

belangrijk zijn voor de studie of functie opgenomen. Dit model wordt gevoed door wetenschappelijke kennis over de relatie tussen studie/werk succes en bepaalde eigenschappen, en door specifieke eisen die aan de studie of functie gesteld worden. Vervolgens kan dit model op verschillende manieren toegepast worden.

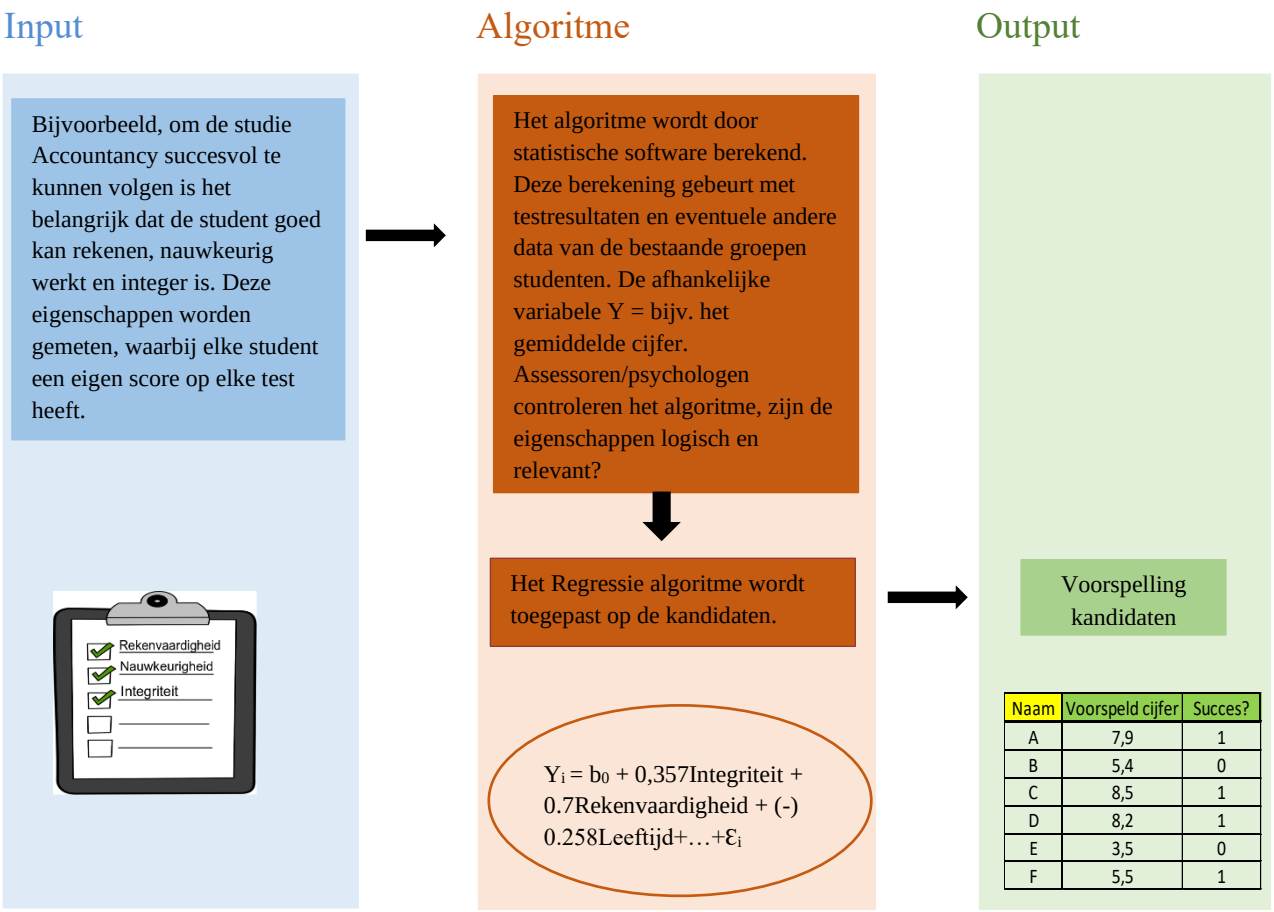
In een klassieke ranking selectie wordt dit model toegepast op de test resultaten van de kandidaten, en het resultaat is een lijst met een overzicht van de best scorende tot de slechts scorende kandidaat. Zie figuur 1 voor een voorbeeld.

Figuur 1. Vereenvoudigd voorbeeld van selectie met behulp van een ranking algoritme.



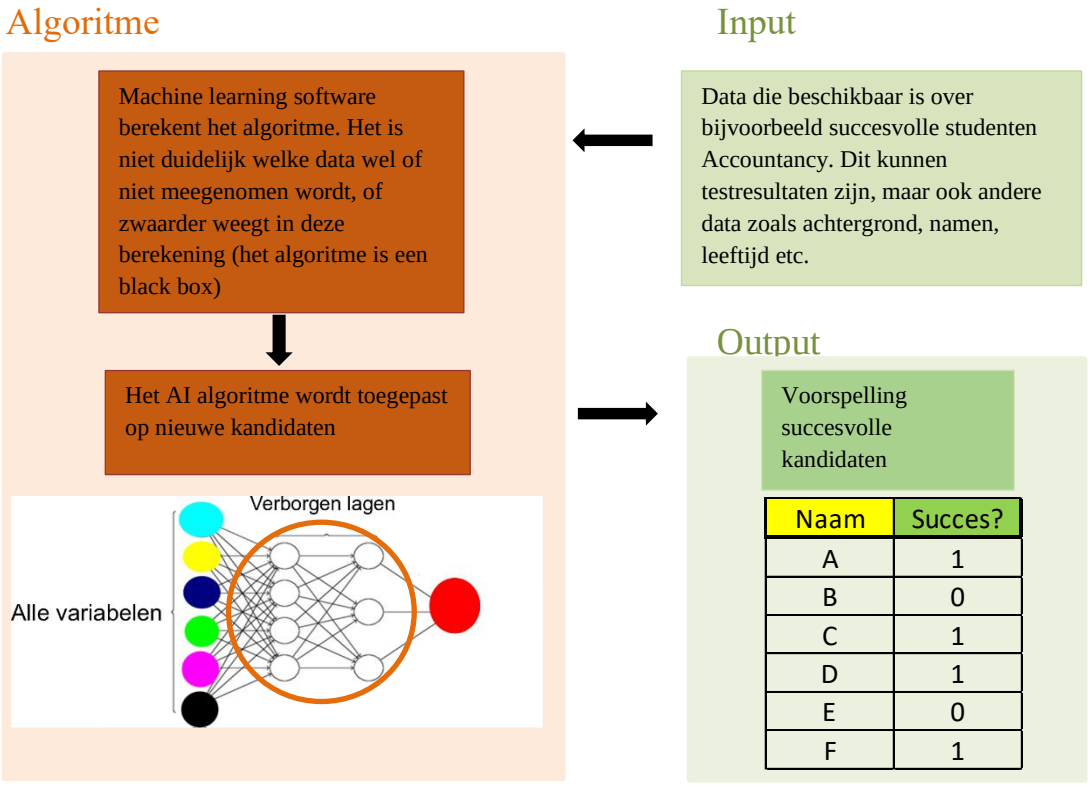
In een selectie met behulp van een regressie model toets je eerst het model met behulp van statistische software op de testresultaten van studenten of werknemers die al studeren of werken. In deze groep zitten de studenten en werknemers waarvan bekend is dat ze wel of niet succesvol functioneren. Deze twee groepen vergelijkt de software op de eigenschappen die je in het model hebt opgenomen. De uitkomst van dit algoritme wordt vervolgens beoordeeld; klopt het model, zijn de belangrijke en relevante eigenschappen aanwezig? Het model kan nu nog aangepast worden. Hierna wordt het model toegepast op de testresultaten van nieuwe kandidaten. In figuur 2 staat hetzelfde voorbeeld als in figuur 1, maar nu uitgewerkt voor het regressiemodel.

Figuur 2. Vereenvoudigd voorbeeld van selectie met behulp van een regressie algoritme.



Sinds enige tijd klinkt ook de roep om bij selectie gebruik te maken van AI. AI is de verzameling software die zelf algoritmes ontwerpt en aanpast. Dit gebeurt bijvoorbeeld door machine learning; algoritmes die ontstaan door patronen in bestaande data te ontdekken, en deze toe te passen op nieuwe data om voorspellingen te doen. Bij AI is er dus geen model wat van te voren ontworpen wordt met behulp van een theoretisch kader, maar dit model wordt gebouwd door de software zelf, vanuit de data van studenten of werknemers die al studeren of werken. Daarmee bouwt de software zelf een model wat toegepast kan worden op nieuwe kandidaten. Alle data die beschikbaar is kan worden opgenomen in de machine learning berekening. De patronen die gevonden worden (het algoritme) blijft onbekend. Zie figuur 3 voor een voorbeeld.

Figuur 3. Vereenvoudigd voorbeeld van selectie met behulp van een AI algoritme.



Het inzetten van AI zou het selectie-en recruitmentproces kunnen verbeteren; de algoritmes van AI zijn snel en efficiënt. Bovendien suggereert de inzet van AI een objectieve en neutrale beslissing: de beslissing om iemand wel of niet te selecteren is het resultaat van een algoritme dat de data neutraal behandelt (Caplan, Donovan, Hanson, en Matthews, 2018). Voorstanders van het gebruik van AI bij recruitment en selectie wijzen er dan ook op dat het kan leiden tot een meer inclusieve werkomgeving. Echter, uit andere publicaties blijkt dat deze beslissingen niet objectief en neutraal zijn, en dat bias op zeer verschillende manieren invloed heeft op de beslissing (zie o.a. Kroll e.a., 2016). De uitkomst kan zijn dat sommige AI algoritmes groepen mensen bevoordelen en anderen benadelen (o.a. O'Neil, 2016). Beslissingen genomen met behulp van AI zijn dus niet fundamenteel meer logisch of eerlijker dan beslissingen door mensen (Caplan e.a., 2018).

2. Opzet van het onderzoek

In dit onderzoek vergelijken wij de verschillen in uitkomsten (wel/niet geselecteerd) tussen drie selectiemethoden. Om deze methoden te vergelijken is een dataset nodig waarin zowel alle eigenschappen van kandidaten aanwezig zijn, als ook een uitkomst variabele: succes. De dataset die we gebruiken bestaat uit de selectie van studenten en hun eerste jaar studieresultaten. Het gaat dus over de selectie van studenten, en specifiek om het verschil in selectie van kandidaten als zij beoordeeld worden met behulp van:

A) Ranking model
B) Regressie model
C) AI model.

Voor elk model is het beste model ontwikkeld. Bij model A is dit een ranking gebaseerd op benodigde kwalificaties ingegeven door theoretische overwegingen en eisen van de studie. Model B is een regressie model, een combinatie van de overwegingen van Model A, en model selectie/model evaluatie gebaseerd op de beste voorspellers voor succes. Bij model C wordt het beste AI model gezocht gebaseerd op selectie/model evaluatie.

Vervolgens zijn de geselecteerde personen van de drie modellen vergeleken; verschillen zij op kenmerken zoals achtergrond, gender, en andere diversiteitskenmerken?

Data

Kandidaten

De studenten uit de cohorten 2013-2017 van een deeltijd universitaire opleiding vormen de onderzochte groep. In totaal zijn van 711 studenten gegevens beschikbaar van zowel de studieresultaten en eigenschappen (testinstrumenten en persoonlijke gegevens). De gemiddelde leeftijd bij aanvang van de studie voor deze groep is 26 jaar (M = 26,00, SD = 4,6) en varieert tussen de 19 en 61 jaar, 64 % was man en 36% vrouw.

Variabelen

De capaciteiten, persoonlijkheid, motivatie en competenties van de studenten zijn gemeten tijdens de aanmeld procedure voor de opleiding, dus voordat zij met de studie zijn begonnen. Er zijn verschillende NOA testinstrumenten gebruikt, zie hiervoor Bijlage I.
Voor elke kandidaat waren er 31 variabelen beschikbaar. Van de NOA testinstrumenten: Capaciteiten (5), Competenties (8), Persoonlijkheid (9), Motivatie (4), en daarnaast nog Persoonlijk (5). De persoonlijke variabelen bestaan uit: Vooropleiding (HBO, BSc, MSc), Gender (Man, Vrouw), Achtergrond (Nederlandse geboorte achtergrond, Niet westerse geboorte achtergrond, en Westerse geboorte achtergrond), Aantal uren beschikbaar voor de studie, en Leeftijd.
Van de 711 studenten heeft 51,9% een HBO opleiding, 10,4% een BSc opleiding en 37,7% een MSc opleiding. Voor deze drie groepen studenten geldt dat zij een ander traject volgen in de universitaire opleiding. 83,4% heeft een Nederlandse geboorte achtergrond, 13,5% heeft een Niet westerse geboorte achtergrond, en 3,1% heeft een Westerse geboorte achtergrond.

Uitkomst variabele: Studiesucces

Het ’gemiddelde cijfer aan het eind van collegejaar 1’ is de maat voor succes. Hoe hoger het gemiddelde cijfer, hoe succesvoller de student[1]. Het gemiddelde cijfer aan het eind van collegejaar 1 is per student berekend; het laagste gemiddelde cijfer aan het eind van het eerste

[1] Deze is tot stand gekomen na een overleg met de opleiding waarin mogelijke alternatieven, zoals nominaal studietempo, aantal te volgen en behaalde vakken, of bijvoorbeeld studiepunten, besproken zijn.

collegejaar is een 0, het hoogste is een 8,75. Gemiddeld hebben de studenten een afgeronde 6 behaald (*M* = 5,82, *SD* =1,57).

Daarnaast moet er een afkappunt gedefinieerd worden: waar ligt de grens voor een wel/niet succesvolle student? Welke student wordt geselecteerd? Voor dit onderzoek is een student geclassificeerd als niet succesvol bij een (voorspeld) cijfer lager dan een 5,5, en als succesvol bij een (voorspeld) cijfer van een 5,5 of hoger. Van de 711 studenten heeft 74% succes en 26% geen succes bij deze definitie.

Modellen

Voor Model A en Model B is geselecteerd welke variabelen gebruikt worden in de analyses. Het doel was om voor Model A de beste theoretische verklarende variabelen te vinden, voor Model B idem, maar met daarbij de mogelijkheid om het model aan te passen na de analyses. Voor Model C werden alle variabelen gebruikt.

Model A: Ranking

Voor Model A is door twee psychologen die niet betrokken waren bij het onderzoek een selectie van variabelen gemaakt en een weging samengesteld, gebaseerd op theoretische overwegingen en de eisen van de opleiding. Na overleg is er consensus bereikt over deze weging, het ranking algoritme.

Model B: *Regressie*

Voor Model B zijn eerst de correlaties van alle variabelen, inclusief de interacties tussen alle variabelen, met het gemiddelde cijfer van jaar 1 vastgesteld. Op deze manier is geïnventariseerd welke variabelen een lineaire relatie met het gemiddelde cijfer hebben. Daarnaast hebben we zo inzicht in welke interacties tussen variabelen van invloed kunnen zijn op studiesucces. De data is gesplitst in een test en hold-out set (50/50), met behulp van een random sample methode. Het best passende model is gekozen, wat ook inhoudelijk relevant is, met als afhankelijke variabele studiesucces (‘het gemiddelde cijfer aan het eind van collegejaar 1’), en als onafhankelijke variabelen de capaciteiten, persoonlijkheid, motivatie, competenties en persoonlijke variabelen van de studenten. In het gereduceerde model, waar alleen nog de beste voorspellers aanwezig zijn, is de correlatie tussen de capaciteiten, persoonlijkheid, motivatie, en competenties, met het gemiddelde cijfer aan het einde van collegejaar 1 significant bij de testdata ($r = .36$, $p < .001$)[2]. Bij de hold-out data is de correlatie tussen de capaciteiten, persoonlijkheid, motivatie, en competenties, met het gemiddelde cijfer aan het einde van collegejaar 1 ook significant ($r = .43$, $p < .001$)[3].

Model C: AI model

Het AI model (machine learning) is ontworpen door 3DUniversum, een expert op dit gebied. Een Multilayer Perceptron (MLP) is gebouwd waarbij rekening is gehouden met het feit dat voor een machine learning model de dataset klein en biased is. Biased slaat hier op het feit dat

[2] $R^2 = .132$, Adjusted $R^2 = .099$, $F(13,354) = 3.979$, $p < .001$).
[3] $R^2 = .186$, Adjusted $R^2 = .155$, $F(13,355) = 6.015$, $p < .001$).

de verdeling succes/geen succes ongelijk is; er zijn meer succesvolle gevallen (74%) dan onsuccesvolle gevallen (26%). Ook het aantal kandidaten (711) is klein te noemen voor deze methode. Alle variabelen zijn in het model opgenomen, en de dataset is gesplitst in een training en testing set, en validation set. In verschillende stappen (m.b.v. Bagging, 2-fold cross validation, Adam optimizer, training op balanced training sets) is het beste model gebouwd. Uiteindelijk is er gekozen voor een Bagging MLP model[4].

Voor alle drie de modellen is daarna een voorspelling gedaan voor studiesucces met behulp van de gecreëerde algoritmes. De succesvolle studenten zouden geselecteerd worden, de niet succesvolle niet.

Resultaten

Voorspellende waarde?

Ten eerste kijken we naar hoe goed alle drie de modellen zijn in het voorspellen van (studie) succes. In figuur 4 staat een overzicht van de voorspelling per model. Te zien is dat het Ranking model 69% correct voorspelt, het Regressie model 74%, en het AI model 71% correct.

Figuur 4. Voorspelling per model.

Voorspelling Model A Ranking					
		Nee	Ja	Totaal	
Succes?	Nee	63	122	185	
	Ja	101	425	526	
	Totaal	164	547	711	0,69

Voorspelling Model B Regressie					
		Nee	Ja	Totaal	
Succes?	Nee	89	96	185	
	Ja	86	440	526	
	Totaal	175	536	711	0,74

Voorspelling Model C AI					
		Nee	Ja	Totaal	
Succes?	Nee	63	122	185	
	Ja	82	444	526	
	Totaal	145	566	711	0,71

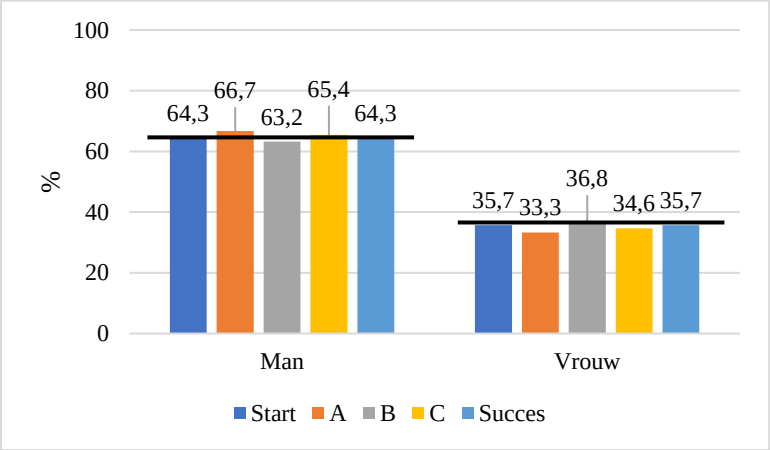
Een eerste conclusie zou dus kunnen zijn dat ten opzichte van de traditionele selectie methoden AI niet veel beter, maar ook niet slechter voorspelt welke student succesvol is. Dit

[4] Voor een volledig verslag van de procedure is een samenvatting van het rapport van 3DUniversum beschikbaar.

sluit aan bij onderzoek wat aangeeft dat een handmatig of regressie algoritme net zo goed of beter is dan een AI algoritme (Narayanan, 2019).

Diversiteitskenmerken: gender en geboorte achtergrond
Vervolgens vergelijken we de modellen op de diversiteit kenmerken van geselecteerden. Aan de start van de studie is 64,3% man en 35,7% vrouw, en aan het einde van jaar 1 zijn er geen verschillen in hoe succesvol deze groepen zijn. Voor Gender lijkt het dat zowel het Ranking model (A) als het AI model (C) het percentage succesvolle vrouwen onderschat (en dus het percentage succesvolle mannen overschat. Het Regressie model (B) lijkt deze bias omgekeerd te hebben, dat model voorspelt een hoger percentage succesvolle vrouwen (zie figuur 5). De verschillen zijn echter niet significant.[5]

Figuur 5. Percentage mannen en vrouwen: Aan de start van de studie, na 1 jaar, en de voorspellingen voor na 1 jaar.



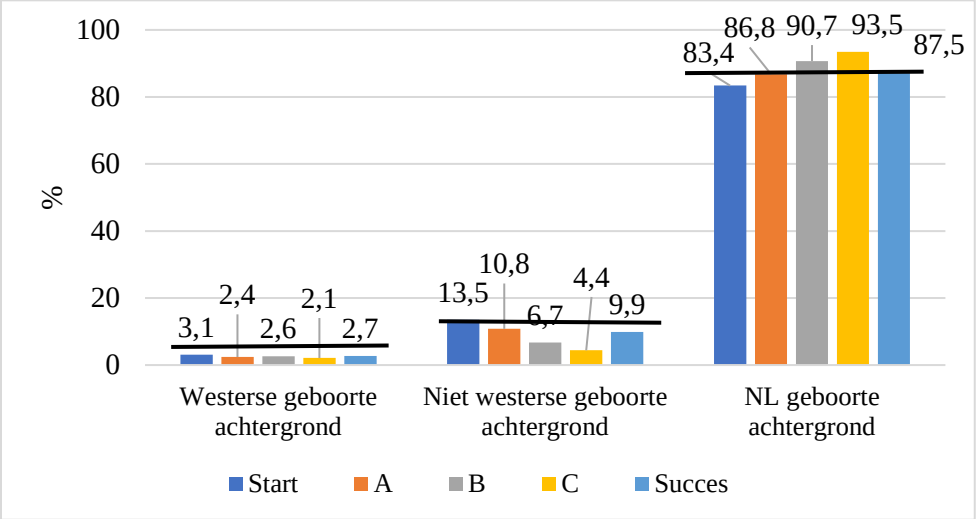
Voor geboorte achtergrond geldt dat er verschillen zijn tussen de groepen aan het begin van de studie, en succes aan het einde van jaar 1 (zie figuur 6). 83,4% van de startende studenten heeft een Nederlandse geboorte achtergrond, en 87,5% van de studenten die succesvol zijn heeft een Nederlandse geboorte achtergrond. 13,5% van de startende studenten heeft een Niet westerse migratieachtergrond, en 9,9% van de succesvolle studenten heeft een Niet westerse migratieachtergrond. De frequenties zijn significant ongelijk, $X(2) = 6{,}45$, $p = .04$.

Deze verschillen worden vergroot in de voorspellingen die de modellen maken.
Alleen het Ranking model (A) heeft geen onder- of overschatting op achtergrond[6]. Het Regressie model (B) en het AI model (C) overschatten beide het succes van studenten met een Nederlandse geboorte achtergrond, ten koste van studenten met name een Niet westerse geboorte achtergrond in hun voorspellingen[7]. Het Regressie model voorspelt dat van de succesvolle studenten 90,7% een Nederlandse achtergrond heeft, en 6,7% een Niet westerse geboorte achtergrond. Het AI model voorspelt dat van de succesvolle studenten 93,5% een

[5] Model A Ranking: $X(1) = 1.40$, $p = .236$, Model B Regressie: $X(1) = .259$, $p = .611$, Model C AI: $X(1) = .283$, $p = .595$
[6] Model A Ranking: $X(2) = 4.68$, $p = .096$
[7] Model B Regressie: $X(2) = 22.08$, $p < .001$, Model C AI: $X(2) = 43.22$, $p < .001$.

Nederlandse geboorte achtergrond heeft, en 4,4% een Niet westerse geboorte achtergrond. Het Regressie model heeft dus een wat gematigder onder-en overschatting vergeleken met het AI model. Een mogelijke verklaring hiervoor kan zijn dat in het AI model geboorte achtergrond ook als variabele was opgenomen, in tegenstelling tot het Regressie model.

Figuur 6. Percentage naar achtergrond: Aan de start van de studie, na 1 jaar, en de voorspellingen voor na 1 jaar.



Conclusie

De vraag was hoe groot de bias van een AI methode voor recruitment en selectie is, en welke groepen aankomend werknemers en studenten nu precies voor- of nadeel ondervinden? We hebben hiervoor naar één groep gekeken; aankomend studenten aan een universitaire deeltijdopleiding. De vergelijking van de drie selectie methoden leidt tot een aantal conclusies. Ten eerste lijken we in dit onderzoek een lichte bias voor gender te zien, maar de verschillen zijn niet significant. Voor geboorte achtergrond vinden we wel een bias bij het Regressie model en het AI model; beide modellen overschatten het succes voor de groep met de Nederlandse geboorte achtergrond, wat vooral ten koste lijkt te gaan voor de voorspelling van het succes van de studenten met een Niet westerse geboorte achtergrond. Het AI model heeft de grootste overschatting. De bestaande bias in de samenstelling van het huidige studenten bestand wordt dus overgenomen, of zelfs versterkt, door het gebruik van AI bij de selectie van nieuwe studenten.

Een limitatie is dat we in dit onderzoek de bias in de methoden met behulp van de data van een heel specifieke groep hebben onderzocht, te weten aankomend studenten van één opleiding. Meer onderzoek, waarbij naast de data van andere studenten ook de data van sollicitanten en werknemers betrokken wordt kan duidelijk maken of, en wanneer, deze bias verder voorkomt. Een tweede limitatie is dat de voorspellende variabelen in onze dataset bestonden uit gevalideerde psychometrische testinstrumenten (naast de persoonlijke data). Men hoopt dat alle selecties met behulp van dergelijke testinstrumenten verlopen, maar niet uit te sluiten is dat in de praktijk ook op andere manieren eigenschappen worden gemeten. De vraag is of gebruik van dergelijke data de bias (verder) beïnvloedt in de praktijk.

Meer algemeen was dit onderzoek ook bedoelt om te onderzoeken of selectie met AI efficiënt en toepasbaar is. Uit een inventarisatie van de gemaakte werkuren voor het ontwerp van alle drie de modellen blijkt dat het Ranking model het minste uren kosten (± 6), gevolgd door het Regressie model (± 16) en daarna het AI model (± 24). In deze fase van een selectie is het AI model dus niet direct kosten effectiever. Ook bleek dat het AI model niet beter (accurater) voorspelde welke student succesvol was dan het Regressie model. Ten slotte bleek dat de gebruikte dataset van 711 kandidaten nogal klein was om een AI model te bouwen. Dit feit is mede debet aan de extra uren die het bouwen gekost heeft. Dit heeft ook consequenties voor de toepasbaarheid; om een goed AI model te maken is er veel trainingsdata nodig, en lijkt dit met name geschikt voor grote bedrijven. Dat zou er toe kunnen leiden dat alleen voor functies waarvoor de data beschikbaar is van enkele duizenden (succesvolle) huidige werknemers een model gemaakt kan worden wat toegepast kan worden op nieuwe sollicitanten.

Ten slotte wordt de toepasbaarheid van een AI model ook beperkt doordat het algoritme een black box is. Het kan niet uitgelegd worden aan een kandidaat op welke combinatie van eigenschappen men afgewezen wordt. Er kan ook niet uitgelegd worden waarom het AI model een bias heeft. Bij de Ranking en Regressie modellen kan beide uitgelegd worden welke combinatie van eigenschappen zorgt voor een afwijzing. Ook kan bij beide modellen achterhaald worden welke variabelen de bias verhogen of verlagen door aan model evaluatie en selectie te doen, met andere woorden door het algoritme aan te passen.

De eindconclusie van dit onderzoek is daarom dat de combinatie van het gebrek aan transparantie van het algoritme, de hier gevonden bias, en het gebrek aan superieure voorspelkracht ten opzichte van de andere hier onderzochte methoden maakt dat selectie met behulp van AI (nog) niet goed toepasbaar is.

Referenties

Caplan, R., Donovan, J., Hanson, L., & Matthews, J. (2018). Algorithmic Accountability: A Primer. *Data & Society*, link: https://datasociety.net/output/algorithmic-accountability-a-primer/.

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633-705.

Narayanan, A. (2019). *How to recognize AI snake oil*. Princeton University, Department of Computer Science

O'Neil, Cathy (2016). *Weapons of Math Destruction*. New York: Crown.

BIJLAGE I

Tabel 1. Overzicht NOA test instrumenten.

Testinstrument	Variabelen
Cognitieve capaciteitentests	Logisch redeneren (Exclusie) Numeriek inzicht (Cijferreeksen) Snelheid en nauwkeurigheid (Controleren) Taalinzicht (Woord-relaties) Verbaal redeneren (Woord-analogieën)
Competentietest	Analyseren Nauwkeurigheid Plannen en organiseren Doorzettingsvermogen Reflecteren Stressbestendigheid Mondeling communiceren Schriftelijk communiceren
Persoonlijkheidstest	Integer studiegedrag Zelfdiscipline Leergierigheid Creativiteit Extravert studiegedrag Studie initiatief Regels en ordelijkheid Vriendelijk en sociaal gedrag Zelfvertrouwen en faalangst
Motivatietest	Intrinsieke Motivatie Extrinsieke Motivatie Prestatie Motivatie Zekerheid en vertrouwen