Wat is de meerwaarde van Artificiële Intelligentie (AI) in statistisch onderzoek?

Lydia Geijtenbeek[1]
CBS
8 juli 2024

[1] Met dank aan Joep Burger en Jan van der Laan voor het delen van hun methodologische expertise en hun aanvullingen op de tekst, Susan van Dijk voor het kritisch aanscherpen van de teksten, en verschillende collega's waaronder Kevin de Groot, Ineke Bijlsma en Manon Joosten voor hun goede aanvullingen.

Samenvatting

In deze notitie beschrijven we mogelijke voordelen van Artificiële Intelligentie[2] (AI) of Machinaal Leren (ML) voor statistisch onderzoek. We beperken ons tot methoden die een bepaalde uitkomst (bv. inkomen, werkloosheid) van een groep in beeld te brengen in relatie tot bepaalde achtergrondkenmerken (bv. geslacht, type woning). We vergelijken vijf methoden, waarvan er drie onder de noemer AI vallen. Van eenvoudig tot complex zijn dat:

1. Kruistabellen
2. Regressieanalyse
3. Eenvoudige beslisboom (AI/ML)
4. Complexe beslisboom (AI/ML)
5. Neurale netwerken (AI/ML)

Van de ene kant, geldt over het algemeen dat complexere methoden meer ontwikkel- en rekentijd kosten, en dat de resultaten lastiger te interpreteren zijn. Van de andere kant kunnen complexe methoden beter inzichtelijk maken hoe verschillende kenmerken samen een uitkomst beïnvloeden, en kan je ze gebruiken om (beterschattingen te maken voor een uitkomst aan de hand van achtergrondkenmerken. In de hoofdtekst van dit document wordt dit nader toegelicht, inclusief voorbeelden en lijsten met voor- en nadelen per methode.

Uiteindelijk hangt het vooral af van de vraag en de situatie welke methode het meest geschikt is. Het onderstaande schema geeft voor een aantal mogelijke vragen van de gebruiker en mogelijke kenmerken van de data voor elk van de vijf methoden aan in hoeverre deze heel geschikt (+), een beetje geschikt (±) of minder geschikt (-) is:

Tabel 1 Mate van toepasselijkheid van de vijf methoden afhankelijk van de vraag en achtergrondkenmerken

Wat is de vraag, en welke achtergrondkenmerken neem je mee?	Kruistabel	Regressie-analyse	Eenvoudige beslisboom	Complexe beslisboom	Neuraal netwerk
Snel inzicht in een uitkomst met enkele achtergrondkenmerken die onderling weinig samenhangen.	+	±	-	-	-
Inzicht in het verband tussen een uitkomst en enkele achtergrondkenmerken die onderling samenhangen	±	+	±	±	-
Inzicht in de mate waarin een uitkomst samenhangt met verschillende achtergrondkenmerken	±	+	+	+	-
Inzicht in het verband tussen een uitkomst en een groot aantal (combinaties van) achtergrondkenmerken, of achtergrondkenmerken die onderling samenhangen	-	±	+	+	-
Groepen identificeren met combinaties van kenmerken waarbij de uitkomsten gemiddeld relatief hoog (of juist laag) zijn	-	-	+	-	-
En berekening die voor elke combinatie van kenmerken een uitkomst schat. Dit is vooral interessant voor gevallen waarbij de uitkomst (nog) niet bekend is.	-	±	±	+	+
Een zo scherp mogelijke schatting van de uitkomst op basis van achtergrondkenmerken, waarbij uitlegbaarheid of transparantie geen issue is.	-	-	-	±	+

Onze conclusie is hiermee dat AI/ML zeker nuttig kan zijn, afhankelijk van de vraag en de data. Vooral in gevallen met veel achtergrondkenmerken die bovendien onderling samenhangen, heeft AI/ML vaak meerwaarde.

[2] Er zijn veel verschillende definities van AI, en niet iedereen zou de methodes die hier beschreven worden onder AI scharen. Daarom gebruiken we in dit document bij voorkeur de term Machinaal Leren (ML).

1 Inleiding

1.1 Introductie

Het ministerie van Binnenlandse Zaken en Koninkrijksrelaties (BZK) heeft het CBS gevraagd om aan te geven wat de meerwaarde is van complexere methoden zoals Artificiële Intelligentie (AI) ten opzichte van meer traditionele statistische methoden. AI is echter een breed begrip waar veel onder valt, en met veel verschillende definities. Daarom gebruiken we hier de term Machinaal Leren (ML). In deze notitie richten we ons op rekenmethoden om binnen een bepaalde dataset een *uitkomst* (bv. inkomen, energieverbruik, aantal kinderen) te beschrijven aan de hand van *achtergrondkenmerken* (bv. leeftijd, samenstelling van het huishouden, inkomensbron). We bekijken een aantal methoden waarvan er drie vallen onder de noemer AI/ML[3]. Op volgorde van eenvoudig tot complex zijn dat:

1. Kruistabellen
2. Regressieanalyse
3. Eenvoudige beslisboom (AI/ML)
4. Complexe beslisboom (AI/ML)
5. Neurale netwerken (AI/ML)

Voor elk van deze methoden geven we een korte beschrijving, plus voor- en nadelen en we geven aan in welke situaties ML-methoden een meerwaarde kunnen hebben.

Nota bene:
- Al deze methoden kunnen enkel in beeld brengen *hoe* kenmerken met een bepaalde uitkomst samenhangen, maar niet *waarom*. Het gaat dan ook enkel om verbanden, maar niet over oorzaak en gevolg.
- Er bestaan veel meer methoden dan genoemd in deze notitie. We richten ons hier vooral op methoden die binnen het CBS worden gebruikt.

1.2 Voorbeelden

Om de beschrijving tastbaar te maken gebruiken we voorbeelden. Deze introduceren we hier.

Voorbeeld: kleinkinderen

Eerst een theoretisch voorbeeld, waarbij we kijken naar de uitkomst 'heeft kleinkinderen'. Deze komt vrijwel alleen voor bij een selecte groep, namelijk ouderen (55+).

Voorbeeld: energiearmoede

Het CBS gebruikt beslisbomen om voor gemeenten en BZK onderzoek te doen naar groepen met een grote kans op energiearmoede[4]. Als voorbeeld gebruiken we een *vereenvoudigde versie* van dit onderzoek om beslisbomen te vergelijken met tabellen en regressie. Deze versie kijkt naar 100.000 woningen, en bekijkt enkel de volgende achtergrondkenmerken: type huishouden, leeftijd hoofdbewoner, bouwjaar, woningtype, oppervlakte, type eigendom.

Voorbeeld: kinderarmoede

In opdracht van BZK heeft het CBS de afgelopen jaren onderzoek uitgevoerd naar kenmerken die samenhangen met de kans om uit armoede te komen of arm te blijven[5,6]. Hierbij werd gebruik gemaakt van ML in de vorm van een complex beslisboom algoritme (XGBoost). Voor het onderzoek over kinderarmoede heeft het CBS

[3] Meer specifiek: supervised Machine Learning/Machinaal Leren. Bij supervised ML is er een dataset beschikbaar met daarin voor (een representatieve steekproef van) objecten de achtergrondkenmerken en uitkomst. Aan de hand van deze dataset wordt een model geschat dat zo goed mogelijk de uitkomst voorspelt.

[4] https://www.cbs.nl/nl-nl/maatwerk/2024/24/energiearmoede-voor-gemeenten-2021

[5] https://www.cbs.nl/nl-nl/over-ons/innovatie/project/risicofactoren-voor-transities-in-en-uit-armoede

[6] https://www.cbs.nl/nl-nl/over-ons/onderzoek-en-innovatie/project/risicofactoren-voor-armoede-18-30-en-40-64-jarigen-in-armoede

onderzocht welke kenmerken bijdragen aan de kans dat een arm kind het volgende jaar uit de armoede komt, wat die kans is, en welke combinaties van kenmerken samengaan met een hoge of lage kans.[7]

2 Vergelijking van de methodes

2.1 Kruistabellen

Een kruistabel geeft voor elk deelgroep een statistiek, zoals het aantal, percentage of het gemiddelde. Bij de tabellen moet je vooraf aangeven welke kenmerken uitgesplitst worden en hoe. De meeste publicaties van het CBS bestaan uit kruistabellen op StatLine.

Een kruistabel geeft één resultaat:

1. Welke categorieën van een achtergrondkenmerk samenhangen met een hoge of lage uitkomst.

Voordelen

- Eenvoudig samen te stellen.
- Relatief eenvoudig te begrijpen, mits er slechts enkele kenmerken zijn.
- Je kan makkelijk twee groepen met elkaar vergelijken.

Nadelen

- Bij meerdere kenmerken tegelijk wordt een tabel snel groot en onoverzichtelijk.
- Je ziet niet het hele plaatje: omdat je kenmerken 'los', of in combinatie met slechts enkele andere kenmerken ziet. Hierdoor kan je belangrijke verbanden missen, of verkeerde conclusies trekken. Doordat kenmerken onderling samenhangen, kan een uitkomst die eigenlijk maar met een of twee kenmerken samenhangt ook bij andere kenmerken belangrijk lijken (zie voorbeelden).
- Waardekenmerken (bv. leeftijd, woningoppervlakte) moeten vooraf ingedeeld worden in klassen.

Voorbeeld: kleinkinderen

Of iemand kleinkinderen heeft hangt in de eerste plaats samen met leeftijd. Leeftijd hangt echter samen met veel andere kenmerken; zo ontvangen vaker AOW of huishoudelijke hulp (WMO), en zijn ze vaker vrouw. Dus als je tabellen maakt over 'heeft kleinkinderen', dan zie je niet alleen hoge percentages bij ouderen, maar ook bij AOW-ontvangers, personen met WMO-hulp, of vrouwen. Dit geeft een vertekend beeld, en leidt snel tot verkeerde conclusies. Zo kan iemand die ziet dat personen met WMO-hulp vaker kleinkinderen hebben ten onrechte concluderen dat het krijgen van kleinkinderen personen meer hulpbehoevend maakt. Voor deze uitsplitsingen geldt dat er een verband (correlatie) *lijkt* te zijn dat er niet is.

Voorbeeld: energiearmoede

Huishoudens met energiearmoede hebben een laag inkomen en ofwel een hoog energieverbruik ofwel een slechte woning. Laten we tabellen gebruiken om deze belangrijke vraag te beantwoorden: bij welke woningen komt energiearmoede het meest voor?

Tabel 2 geeft een kruistabel met woningkenmerken. De getallen geven het percentage energiearmoede (met kleurcode: rood=hoog en groen=laag). Hierin is te zien dat energiearmoede meer voorkomt in kleine dan in grote woningen, en meer in huurwoningen dan in koop, en meer in hoekwoningen en meergezinswoningen (appartementen). Dat is opvallend: het energieverbruik is lager voor kleinere woningen en woningen met minder buitenmuren, zodat je bij deze woningen juist minder energiearmoede zou verwachten. Misschien hebben kleine woningen vaker energiearmoede omdat het vaker huurwoningen zijn? Uit deze tabel kan je dit niet opmaken.

Tabel 2 Energiearmoede (%) naar woningkenmerken, kruistabel

Woningtype		Oppervlakte		Type eigendom	
Vrijstaande woning	2,2	2 tot 50 m2	13,1	Woningcorporatie	14,6
Twee-onder-een-kapwoning	4,6	50 tot 75 m2	11,9	Koopwoning	1,1
Hoekwoning	8,2	75 tot 100 m2	8,8	Overige verhuur	13,3

[7] https://www.cbs.nl/nl-nl/longread/aanvullende-statistische-diensten/2024/kenmerken-die-samenhangen-met-de-kans-om-uit-kinderarmoede-te-komen

Tussenwoning	4,9	100 tot 150 m2	4,1
Meergezinswoning	9,1	150 tot 250 m2	1,7
		250 of meer m2	2,4

2.2 Regressieanalyse

Bij een regressieanalyse stel je een formule op waarmee op basis van achtergrondkenmerken de uitkomst geschat wordt. Hierbij staan de gebruikte achtergrondkenmerken en de vorm van de formule vooraf op hoofdlijnen vast, en berekent een algoritme[8] hoe sterk en in welke richting (positief of negatief) een kenmerk met de uitkomst samenhangt. Regressie modellen worden meestal vanuit een eenvoudig model aan de hand van theorie opgebouwd naar een complexer model met steeds meer kenmerken. Hierbij is het ook mogelijk om te toetsen of een extra kenmerk nog iets toevoegt of dat het kenmerk gegeven de al in het model aanwezige kenmerken eigenlijk geen effect heeft.

Een regressie geeft twee resultaten:

1. Per kenmerk of interactieterm een coëfficiënt (getal) die aangeeft hoe en hoe sterk deze samenhangt met de uitkomst.
2. Een formule waarmee je voor elke combinatie van kenmerken de bijbehorende uitkomst kan schatten.

Voordelen

- Kenmerken kunnen in samenhang bekeken worden.
- Minder schijnverbanden. Als verschillende kenmerken met elkaar en de uitkomst samenhangen, dan zal een groter deel van de samenhang worden toegeschreven aan kenmerken met een sterkere relatie met de uitkomst.
- Waardekenmerken (bv. leeftijd of oppervlakte) kunnen als getal meegenomen worden (al moet je wel aangeven in welke vorm).
- Als er een goede theorie is, is het relatief makkelijk om vanuit de theorie een model op te bouwen. Ook kan je dan goed toetsen of het model inderdaad voldoet aan de verwachtingen.

Nadelen

- Soms ingewikkelder te interpreteren dan tabellen. Vaak is een rapport nodig in plaats van enkel een tabel.
- Een regressie is gevoelig voor misspecificatie van het model. Als je aannames doet die niet blijken te kloppen, geven de uitkomsten een vertekend beeld.
- Meer werk aan data voorbereiding en analyse:
    i. Regressie kan meestal[9] niet goed overweg met grote aantallen kenmerken; de benodigde rekenkracht en geheugen nemen dan sterk toe.
    ii. Tegelijkertijd is het belangrijk dat alle belangrijke kenmerken worden meegenomen, omdat je anders (net als bij tabellen) gemakkelijk verkeerde conclusies trekt.
    iii. Voor alle kenmerken moet vooraf de vorm bepaald worden, bv: indeling in klassen, lineair, kwadratisch.
    iv. Combinaties van kenmerken worden alleen in samenhang meegenomen als dat expliciet (in de vorm van een interactie) aan het model meegegeven is.
- Als variabelen sterk samenhangen (bv. leeftijd en werkervaring, of inkomen en loon), dan geven de geschatte coëfficiënten mogelijk een vertekend beeld. Dit geldt ook voor modellen waarin verschillende combinaties van kenmerken zijn opgenomen.
- Risico dat relevante kenmerken over het hoofd worden gezien omdat ze niet geselecteerd zijn of niet in de goede vorm in het model zitten.

Voorbeeld: kleinkinderen

Terug naar het theoretische voorbeeld over de uitkomst ‘heeft kleinkinderen’. Omdat in een regressie alle variabelen tegelijkertijd beschouwd worden, zie je bij een regressie nog steeds terug dat leeftijd heel belangrijk is, maar variabelen die niks met kleinkinderen te maken hebben en wel met leeftijd blijken dan niet of amper relevant. Voor sommige variabelen zal het verband zelfs omdraaien: omdat ouderen vaker alleen wonen zie je

[8] Veelgebruikte algoritmen voor regressie zijn OLS (Ordinary Least Squares), of optimalisatie van ML (Maximum Likelihood).
[9] Je kan dit eventueel oplossen, bijvoorbeeld via Principal Component Analysis. Maar daarmee wordt je regressie wel weer een stuk complexer en daarmee moeilijker te interpreteren

in een kruistabel dat eenpersoonshuishoudens vaker kleinkinderen hebben. In een regressie vergelijk je typen huishouden van dezelfde leeftijd, en blijkt dat (oudere) paren juist vaker kleinkinderen hebben dan (oudere) alleenstaanden.

Voorbeeld: energiearmoede
In Tabel 2 zagen we hoe verschillende woningtypen samenhangen met energiearmoede. Maar omdat deze onderling ook weer samenhangen (bv. grote woningen zijn vaker koop, kleine woningen zijn vaker een appartement), was het niet duidelijk welke verbanden sterker zijn, en welke afgeleid. Regressie kan hierbij helpen.

In Tabel 3 staan de uitkomsten een regressie met meerdere kenmerken[10] naast die van een kruistabel. De getallen geven het geschatte effect van een bepaald kenmerk, gecorrigeerd voor de andere achtergrondkenmerken. Deze geven een beter beeld van het echte verband tussen kenmerk en uitkomst. Opvallende verschillen zien tussen kruistabel en regressie:

- In de kruistabel waren er grote verschillen in energiearmoede tussen kleine en grote woningen, maar bij de regressie vallen die bijna helemaal weg. Een reden hiervoor kan zijn dat kleine woningen vaker huurwoningen zijn en personen in een huurwoning vaker last hebben van energiearmoede.
- In de kruistabel hebben vrijstaande woningen weinig energiearmoede en meergezinswoningen veel, maar dit verschil valt weg in de regressie. Misschien komt dit doordat appartementen vaker klein en huur zijn, en vrijstaande woningen eerder koop.

Conclusie: in de kruistabel lijken woningoppervlakte en eigendom even belangrijk, maar de regressieanalyse laat zien dat energiearmoede vooral een probleem is bij huurwoningen en hoek of 2-onder-1 kapwoningen, en dat woningoppervlakte los van eigendom of woningtype nauwelijks invloed heeft.

Tabel 3 Energiearmoede (%) naar woningkenmerken, kruistabel versus regressieanalyse

Woningtype	Tabel	Regressie	Oppervlakte	Tabel	Regressie	Type eigendom	Tabel	Regressie
Vrijstaande woning	2,2	7,3	2 tot 50 m2	13,1	8,4	Woningcorporatie	14,6	7,1
Twee-onder-een-kapwoning	4,6	8,2	50 tot 75 m2	11,9	7,7	Koopwoning	1,1	-5,7
Hoekwoning	8,2	9,7	75 tot 100 m2	8,8	6,6	Overige verhuur	13,3	6,7
Tussenwoning	4,9	6,6	100 tot 150 m2	4,1	6,6			
Meergezinswoning	9,1	4,2	150 tot 250 m2	1,7	7,1			
			250 of meer m2	2,4	6,7			

2.3 Eenvoudige of complexe beslisboom (AI/ML)

Een beslisboom verdeelt eenheden (bv. personen of woningen) op basis van kenmerken in groepen met vergelijkbare uitkomsten binnen elke groep. Een beslisboomalgoritme berekent welke kenmerken en welke klassen of waarden gebruikt worden om de groepen zo goed mogelijk te splitsen.

Er zijn verschillende methoden die gebruik maken van beslisbomen. De meest eenvoudige maakt één boom. Het voordeel van een enkelvoudige beslisboom is dat het eindresultaat eenvoudig te visualiseren als boomdiagram. De kwaliteit van de voorspelling is echter vaak beter als je meerdere bomen gebruikt, zoals bij een Boosted Tree of Random Forest. Een Random Forest bestaat uit een verzameling bomen en is daardoor minder gevoelig voor kleine veranderingen in de waarnemingen. Bij Boosted Trees verbetert iedere volgende beslisboom de afwijkingen in voorspellingen van de voorgaande beslisbomen, waardoor deze als het ware 'leert'.

Een beslisboom geeft verschillende resultaten:
1. Per kenmerk hoe sterk deze samenhangt met de uitkomst (bv. variable importance of SHAP value).
2. Een formule waarmee je voor elke combinatie van kenmerken de bijbehorende uitkomst kan schatten.

[10] De volgende kenmerken zijn meegenomen: type huishouden, leeftijd hoofdbewoner, bouwjaar, woningtype, oppervlakte, type eigendom, allen als categoriale kenmerken en zonder interacties. Verder gaat het om een lineaire regressie (en geen logistische regressie), omdat het daarbij eenvoudiger is om van de uitkomsten een leesbare tabel te maken.

3. *Voor eenvoudige bomen:* een daadwerkelijke beslisboom inclusief omschrijving van groepen met een hoge of lage uitkomst.

Voordelen

- Achtergrondkenmerken worden in samenhang bekeken, ook als deze niet expliciet benoemd zijn.
- Zelfselectie en veel kenmerken. Je hoeft niet vooraf kenmerken te selecteren; in het model kunnen grote aantallen kenmerken meegenomen worden en de methode bepaalt welke belangrijk zijn.
- Complexe fenomenen waarbij veel kenmerken een rol spelen en kenmerken onderling samenhangen kunnen relatief goed beschreven worden. Het is daarbij niet nodig om vooraf al aannames te doen over relaties tussen kenmerken en de uitkomst:
    i. Voor waarde kenmerken (bv. leeftijd of oppervlakte) geldt dat je niet vooraf de vorm hoeft te bepalen (bv. een indeling in categorieën, en of een verband lineair of kwadratisch is).
    ii. Relevante interacties worden (meestal) automatisch meegenomen in het model. Als bijvoorbeeld het effect van leeftijd anders is voor mannen dan voor vrouwen, dan ziet de deelboom voor vrouwen er anders uit.
- Ontbrekende waarden. Beslisbomen kunnen goed omgaan met ontbrekende waarden in kenmerken, zodat je ook kenmerken mee kan nemen die voor een deel van de populatie onbekend zijn.
- *Bij eenvoudige boom*: je kan de boom als output opleveren. Beslisbomen zijn bij steeds meer mensen bekend, en daarmee goed te begrijpen.

Nadelen

- De uitkomsten zijn lastiger te interpreteren:
    i. *Bij eenvoudige bomen* bestaan de uiteindelijke groepen vaak uit combinaties van veel verschillende kenmerken; het is dan lastig om een groep in één zin te omschrijven. Ook ontstaan vaak restgroepen waarvan een groot deel van de uitleg is dat het *niet* een andere groep is.
    ii. *Bij complexe beslisbomen* is het is lastig om te zien *hoe* verschillende variabelen met elkaar samenhangen, en waarom een model bepaalde voorspellingen doet. Daardoor is het ook lastig om te controleren of alles goed gaat.
- Een beslisboom heeft een voorkeur voor variabelen die uit veel verschillende klassen bestaan. Hierbij bestaat echter het risico op *overfitting*, waarbij het model toevallige variaties in uitkomst aanziet voor echte verbanden. Daarom is het vaak nodig om klassen in te dikken.
- Je kan weliswaar veel variabelen meenemen, maar hoe meer variabelen je meeneemt, des te meer tijd is nodig voor datapreparatie en voor het schatten van het model.
- Geen lineaire verbanden. Als de uitkomst en een kenmerk beiden lineair zijn en het verband daartussen is dat ook (bijvoorbeeld energieverbruik en woningoppervlakte), dan kan een beslisboom hier minder goed mee omgaan.[11] Het model wordt dan onnodig groot, terwijl de schattingen mogelijk slechter zijn dan bij een regressie met lineaire term of een neuraal netwerk.
- *Bij eenvoudige boom*: de kwaliteit van de boom hangt sterk af de keuze van achtergrondkenmerken. Als deze niet goed gekozen of ingedeeld zijn, kan de boom een vertekend beeld geven of kan de uitkomst niet goed voorspeld worden.

Voorbeeld: energiearmoede

Een belangrijk voordeel van beslisbomen, is dat je groepen kan afleiden met een hoge of juist lage uitkomst. Dit kan een beetje bij tabellen, maar dan op basis van een of enkele kenmerken. In een regressie kan je losse kenmerken selecteren, maar is het lastiger om die te combineren.

Zie Tabel 4 voor een vergelijking tussen de methoden. In de kruistabel waren de groepen met het grootste aandeel energiearmoede woningcorporaties (14%), overige huurwoningen (13%) en kleine woningen tot 50m2 (13%). Bij een regressie komt er voor elk kenmerk uit welke categorie het meest energiearmoede heeft. Bij een

[11] Bij een lineair kenmerk bepaalt een beslisboom steeds een drempelwaarde waarboven de uitkomst relatief laag is en waaronder relatief hoog. Bij een sterk lineair verband met de uitkomst komt een kenmerk vaak terug met veel drempelwaardes, maar zijn de schattingen van de uitkomst tussen twee drempelwaardes gelijk.

lineaire regressie kan je die bij elkaar optellen, en de som geeft 26,5%. De beslisboom maakt daarentegen simpelweg vijf groepen, waarvan sommige heel weinig energiearmoede hebben en sommige juist veel (zie

Een volgend voordeel van een beslisboom is dat deze behalve een hoge of lage groep, ook de rest van de data opdeelt in groepen met vergelijkbare uitkomsten. Tabel 5 toont alle vijf de groepen die de beslisboomanalyse in het voorbeeld opleverde.

Tabel 5). De groep met het hoogste aandeel energiearmoede is de groep van “huurwoningen voor 1983, eenpersoons/eenouderhuishouden, hoek of 2-onder-1 kap”, met maar liefst 35,5% energiearmoede. Van de drie methoden geeft de beslisboom dus de groep met het hoogste percentage energiearmoede.

In de vergelijking tussen deze regressie en beslisboom valt op dat de beste groep uit de beslisboom niet alleen een hoger percentage energiearmoede heeft, maar ook uit veel meer woningen bestaat (2.300 bij de beslisboom ten opzichte van 1 woning in de regressie), en dat de beslisboom minder kenmerken gebruikt waardoor deze makkelijker te omschrijven is.

Tabel 4 Groep met hoogste aandeel energiearmoede per methode

Methode	Hoogste groep	Energiearmoede (%)
Kruistabel	Verhuur door woningcorporatie	14,6
Regressie	Eenpersoonshuishouden, leeftijd hoofdbewoner 25-45 jaar, hoekwoning, bouwjaar voor 1946, oppervlakte tot 50m^2, corporatiewoning	26,5
Beslisboom	Huurwoning voor 1983, eenpersoons-/eenouderhuishouden, hoek of 2-onder-1 kap	35,5

Een volgend voordeel van een beslisboom is dat deze behalve een hoge of lage groep, ook de rest van de data opdeelt in groepen met vergelijkbare uitkomsten. Tabel 5 toont alle vijf de groepen die de beslisboomanalyse in het voorbeeld opleverde.

Tabel 5 Groepen met een hoge of juist lage kans op energiearmoede

Combinatie van kenmerken	Energiearmoede (%)
Koopwoning	1,1
Huurwoning na 1982	4,7
Huurwoning voor 1983, paar of overig huishouden	12,7
Huurwoning voor 1983, eenpersoons-/eenouderhuishouden, hoek of 2-onder-1 kap	35,5
Huurwoning voor 1983, eenpersoons-/eenouderhuishouden, geen hoek of 2-onder-1 kap	22,0

Een derde voordeel van een beslisboom (maar ook van regressie) is de voorspelkracht.

Je kan regressie of een beslisboom gebruiken om de uitkomst te schatten. Dit kan ook met kruistabellen: het gemiddelde van een groep is een ruwe voorspelling voor de gevallen binnen de groep. Een standaardmaat voor voorspelkracht is de verklaarde variantie (R^2), ofwel het deel van de verschillen in uitkomst dat verklaard kan worden door het model; een model dat alle uitkomsten perfect voorspelt heeft een voorspelkracht van 1, en een model dat er niks van bakt heeft een 0.

Voor zowel de kruistabellen, de regressie als de beslisboom hebben we een R^2 berekend. In alle drie de gevallen zijn de volgende kenmerken meegenomen: type huishouden, leeftijd hoofdbewoner, en van de woning type, bouwperiode, oppervlakte en eigendom. Tabel 6 toon de verklaarde variantie (R^2) voor elk van de methoden[12]. De verklaarde variantie gaat van 0,070 bij kruistabellen tot 0,098 bij regressie en 0,123 bij de eenvoudige beslisboom. Dat is niet zo hoog; dat komt mede doordat in dit voorbeeld sommige belangrijke kenmerken (bv. het energielabel) niet zijn meegenomen. Er is echter wel duidelijk dat in dit voorbeeld de kruistabel de laagste voorspelkracht heeft en de beslisboom de hoogste[13].

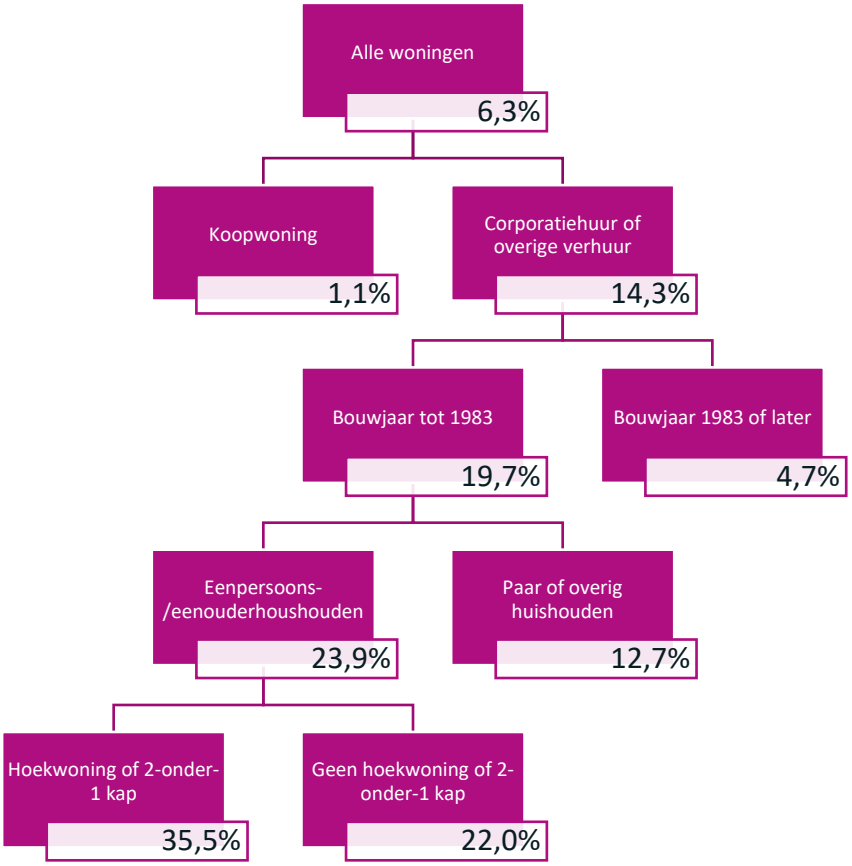
[12] Voor de kruistabel is per tabel de verklaarde variantie berekend, en daarvan de hoogste genomen.
[13] Merk op dat het niet zo is dat een beslisboom altijd beter voorspeld dan bijvoorbeeld een regressiemodel. Een goed gespecificeerd regressiemodel (een model dat goed bij de data past) kan (veel) betere voorspellingen opleveren dan een beslisboom.

Tabel 6 Verklaarde variantie per methode voor energiearmoede met dezelfde achtergrondkenmerken

Methode	Verklaarde variantie (R^2) energiearmoede
Kruistabel, alleen enkelvoudige tabellen	0,070
Regressie, zonder interacties	0,098
Beslisboom, eenvoudig	0,123

Een laatste voordeel van een beslisboom is dat je de uitkomst ook kan laten zien in de vorm van een boomdiagram. zie Figuur 1 voor de boomstructuur voor het voorbeeld energiearmoede. Je kan deze boom doorlopen voor een willekeurige woning door bovenaan te beginnen, en bij elke vertakking de tak te kiezen die van toepassing is.

Figuur 1 Grafische weergave beslisboom energiearmoede, inclusief energiearmoede (%) per groep



Voorbeeld: kinderarmoede

In dit onderzoek is op basis van een groot aantal kenmerken de kans geschat dat kinderen uit armoede komen. In dit onderzoek zijn twee methoden gebruikt:

1. Een complex beslisboomalgoritme, en wel *eXtreem Gradient Boosting* (XGBoost). Deze is gebruikt om kenmerken te bepalen die samenhangen met uitstroom, en om voor elk individu een uitstroomkans te schatten. De kenmerken die het meeste samenhangen met armoede zijn: de leeftijd van het kind; of het huishouden in 2019 al inkomensarm was (weinig inkomen had); of er een verandering was in partnerschap van de vader of moeder; de leeftijd van de moeder bij geboorte.
2. Een eenvoudige beslisboom. Deze is gebruikt om groepen af te leiden met een grote (of juist kleine) kans op uitstroom. De groep met de kleinste kans om uit armoede te komen heeft de volgende kenmerken: het kind is jonger dan 14 jaar; het kind leeft in een huishouden dat moet rondkomen van een uitkering; er is al langer sprake van inkomensarmoede; er is geen verandering geweest in partnerschap van de ouder(s) en er is geen ouder/partner weggegaan uit het huishouden.

Je kan beide algoritmen gebruiken om de uitstroomkans te schatten en de voornaamste kenmerken te bepalen die samenhangen met uitstroom. Dat hebben de onderzoekers ook gedaan. Daarbij bleek echter dat de schatting van de complexe boom veel beter was: de simpele boom kon 34% van de verschillen in uitstroom verklaren aan de hand van achtergrondkenmerken, maar bij de complexe boom was dat maar liefst 76%. De voorspelkracht is dus ruim tweemaal zo hoog.

Conclusie: een belangrijk voordeel van complexe beslisbomen ten opzichte van eenvoudige bomen is dat ze over het algemeen ene grotere voorspelkracht hebben.

2.4 Neurale netwerken (AI/ML)

Een veelgebruikte AI/ML-methode is een Neuraal Netwerk (NN). Een neuraal netwerk is voor te stellen als een net, met knopen en verbindingen. Het net heeft een input-laag met knopen die samen de kenmerken coderen (bv. leeftijd of inkomen), een output-laag die de uitkomst codeert (bv. de persoon is arm), en daar tussenin een of meer verborgen lagen. Als een geval wordt aangeboden aan een NN, krijgen eerst de knopen in de input-laag de waardes van de desbetreffende kenmerken. Daarna worden deze doorgegeven aan de volgende laag, waarbij gewichten bepalen hoeveel van elk kenmerk aan welke knoop wordt doorgegeven, en daar gecombineerd zodat elke knoop een waarde krijgt. Ook deze waardes worden weer doorgegeven en gecombineerd. Dit gaat zo door totdat de output-knoop een waarde krijgt. Dit is de voorspelde uitkomst. Een Neuraal Netwerk is daarmee een soort van complexe formule, die uit elke combinatie van kenmerken een uitkomst berekent.

Als je een neuraal netwerk wil gebruiken, bepaal je eerst hoeveel knopen[14] er moeten komen, hoeveel lagen, en op welke manier de knopen met elkaar verbonden zijn. Dit hangt onder andere af van het aantal kenmerken, het aantal uitkomsten, en de complexiteit van de samenhang. Vervolgens wordt het netwerk *getraind*. Daarbij krijgt het netwerk steeds van een of meer gevallen de kenmerken te zien, en leidt daar een uitkomst uit af. Afhankelijk van hoezeer deze samenhangt met de echte uitkomst worden de gewichten in het netwerk aangepast. Dit wordt vele malen herhaald, totdat het netwerk bijna niet meer verandert. Het resultaat is een netwerk dat voor elke combinatie van kenmerken een uitkomst schat.

Voordelen

- Grote voorspelkracht, met name bij complexe verbanden.
- De vorm van de verbanden hoeft niet vooraf meegegeven te worden, en kan complexere vormen aannemen dan bij een beslisboom.
- Extrapoleren. Een neuraal netwerk geeft vaak betere schattingen voor waardes of combinaties van kenmerken die niet voor kwamen in de data waarop het model getraind is.

Nadelen

- Black box. Een neuraal netwerk is niet goed uitlegbaar; door de vele verbanden en complexe berekening is het zeer lastig om de volgen hoe het model tot zijn voorspellingen komt.
- Het risico is hoger dat het model onlogische of ongewenste verbanden schat (bijvoorbeeld of een model discrimineert op basis van herkomst). Zoals in voorgenoemde punt is aangeven.
- Om een neuraal netwerk te schatten is vaak complexere programmatuur nodig, en een computer met veel rekenkracht en geheugen. Ook is het belangrijk dat het aantal knopen en lagen goed gekozen wordt, en de methode om fouten te corrigeren. Dit vergt expertise en tijd. Een neuraal netwerk is daarmee de duurste methode.

3 Vergelijking methoden

In het vorige hoofdstuk hebben we vijf methoden vergeleken, en gekeken naar de voor- en nadelen van deze methoden. Tabel 7 geeft een samenvatting over hoe verschillende methodes scoren op een aantal relevante aspecten. Hieronder volgt een meer algemene beschrijving van een paar belangrijke aspecten waarop deze methoden en welke methode in bepaalde gevallen geschikt is:

[14] Waarbij het aantal knopen in de input-laag afhangt van de kenmerken die je meeneemt en hoe je die meeneemt, en het aantal knopen van de output-laag afhangt van de doelvariabele(n).

- Doel. Waar wil je statistiek voor gebruiken? Wil je het verband weten tussen de uitkomst en een categoriaal (gemaakt) achtergrondkenmerk? Neem dan een kruistabel. Wil je weten welke kenmerken het sterkste samenhangen met de uitkomst? Gebruik dan regressie of een beslisboom. Wil je groepen maken op basis van combinaties van kenmerken met hele hoge of lage uitkomst? Neem een eenvoudige beslisboom. Of wil je aan de hand van kenmerken de kans schatten op een bepaalde uitkomst? Dan kan je regressie, een beslisboom of een neuraal netwerk gebruiken.
- Begrijpelijkheid. Hoe lastig is het voor een gebruiker van de data om de uitkomsten te lezen en interpreteren? Een kruistabel met een of twee variabelen is voor veel mensen goed te begrijpen, en een groep uit een beslisboom die bestaat uit 2-5 kenmerken ook. De geschatte coëfficiënten van een regressie, een score voor het belang van variabelen of de boomdiagram van een eenvoudige beslisboom zijn voor de meeste mensen ook wel te volgen. De formule van een regressie is al iets ingewikkelder. En een complexe beslisboom of neuraal netwerk valt eigenlijk niet direct te lezen. Er zijn wel methoden die helpen bij de interpretatie van dit soort modellen.
- Kenmerken. Waar moet je op letten bij de achtergrondkenmerken? Bij beslisbomen of neurale netwerken kan je veel kenmerken meenemen, bij kruistabellen of regressie minder. Met name bij regressie is er vaak een uitgebreide voorbewerking van de kenmerken nodig, en moet je goed opletten welke variabele mee kunnen en zo ja in welke vorm. Ook bij kruistabellen moet je vooraf nadenken over een indeling. AI/ML-methoden zijn flexibeler in de hoeveelheid en vorm van de variabelen.
- Complexiteit, ofwel hoe ingewikkeld is het om een methode toe te passen? Complexere methoden vergen meer van de onderzoeker, kosten meestal meer tijd, en vragen vaak ook meer rekenkracht. Een kruistabel maken is vrij eenvoudig. Een regressie of eenvoudige beslisboom is iets ingewikkelder, je moet vaak dingen testen en controleren, en enige wiskundige kennis is vereist. Voor een complexe beslisboom moet je een groot aantal instellingen bepalen en testen, op verschillende manieren de kwaliteit in de gaten houden, en vereist dat je kan programmeren en een gedegen wiskundige of statistische basis hebt.

Tabel 7 Samenvatting van voor- en nadelen van methoden

Aspect van de methodes	**Kruistabel**	**Regressie**	**Eenvoudige beslisboom**	**Complexe beslisboom**	**Neuraal netwerk**
Begrijpelijkheid van het eindresultaat	Hoog, mits 1-2 variabelen	Midden	Hoog/Midden	Midden/Laag	Laag
Complexiteit van methode	Laag	Laag/Midden	Midden	Midden/Hoog	Hoog
Voorspelkracht van het model	Laag	Midden	Midden	Midden/Hoog	Midden/Hoog
Hoe leidt je het belang van kenmerken af?	Via verschillen in uitkomst (X^2 of RL toets)	Via p-/T-waarde	Via maat voor 'importance'	Via maat voor belang bij schatting (bv. SHAP)	ia maat voor belang bij schatting (bv. SHAP)
Kan je groepen afleiden op basis van enkele kenmerken met hoge/lage uitkomst?	Enkel per kenmerk	Enkel per kenmerk	Ja	Nee	Nee
Hoeveelheid kenmerken kan je (tegelijkertijd) meenemen?	Enkele (1-3)	Niet te veel (5-20)	Mogen er veel zijn (>100)	Mogen er veel zijn (>100)	Mogen er veel zijn (>100)
Beschouwt de methode de kenmerken los van elkaar of in samenhang?	Losse kenmerken	Vooral losse kenmerken, beetje samenhang	Lokale samenhang deel binnen een tak van de boom	Samenhang binnen delen van de boom en tussen bomen	Volledige samenhang
Kan omgaan met interacties of non-lineaire verbanden?	Nee	Beetje, mits je die er vooraf instopt	Ja	Ja	Ja
Kosten (tijd)?	Laag	Midden	Midden	Hoog	Zeer hoog

4 Conclusie

In deze notitie zijn verschillende methoden voor statistisch onderzoek vergeleken, waaronder AI/ML-methoden. Elke methode heeft voor- en nadelen en de voornaamste conclusie is dat het van de data en de doelen afhangt welke methode het meest geschikt is.

Over het algemeen geldt dat een complexe (ML) methode vaak kwalitatief betere uitkomsten geeft. Het nadeel is dat deze wel meer tijd kost en lastiger te begrijpen is. Overigens geldt dit zowel tussen methoden als binnen één enkele methode: een ML-model geeft vaak een betere schatting dan een regressie, maar een regressie met veel interactietermen geeft ook vaak een betere schatting dan een eenvoudige regressie. Daarnaast hangt het van de methode af wat je er wel en niet mee kan: met een kruistabel kan je niet bepalen welke kenmerken het meest relevant zijn voor de uitkomst, en kan je ook niet een uitkomst schatten op basis van achtergrondkenmerken. Daarvoor zijn complexere methoden nodig zoals een regressie, of ML-methoden zoals een beslisboom of neuraal netwerk.

Wat de beste methode is hangt vooral af van de soort data en het doel van het onderzoek:

- Als er maar weinig kenmerken samenhangen met de uitkomst, en de samenhang is bovendien simpel en eenduidig, gebruik dan kruistabellen. Andere methoden voegen weinig toe, maar kosten wel meer tijd en zijn minder eenvoudig te begrijpen.
- Als je wil weten voor welke combinaties van kenmerken de uitkomst hoog of juist laag is, gebruik dan kruistabellen of eenvoudige beslisbomen. Bij meer dan 2 of 3 kenmerken wordt een kruistabel onoverzichtelijk, en werkt een beslisboom meestal beter.
- Als je wil weten welke kenmerken het meeste samenhangen met een hoge of lage uitkomst, gebruik dan regressie of een beslisboom. Regressie is relatief snel en eenvoudig, maar werkt alleen goed als er niet te veel kenmerken zijn, en als vooraf duidelijk is hoe de kenmerken met de uitkomst samenhangen. Een beslisboom kan omgaan met een groot aantal kenmerken, en met kenmerken die onderling samenhangen. Een complexe beslisboom kost meer tijd, maar geeft ook betere resultaten.
- Als je een schatting wilt van de uitkomst op basis van de kenmerken, gebruik dan regressie, een beslisboom of een neuraal netwerk. Je gebruikt een regressie als er niet te veel kenmerken zijn, als de onderlinge samenhang beperkt is, of als vooraf duidelijk is hoe de uitkomst met de kenmerken samenhangt. Als er veel kenmerken zijn,

Tot slot: het kan het zinvol zijn om methoden te combineren. Zo kan je een eenvoudige boom gebruiken om te bepalen welke (combinaties van) kenmerken relevant zijn, en vervolgens een kruistabel of regressie maken met alleen de meeste relevante kenmerken en combinaties daarvan. Of net als bij kinderarmoede een complexe beslisboom gebruiken om af te leiden wat de belangrijkste kenmerken zijn, en daarnaast een eenvoudige beslisboom om groepen af te leiden met een hoge kans op armoede.