Het voorspellen van voortijdig schoolverlaters

Een Machine Learning onderzoek

M.A. Kuijer

1764608

13 augustus 2015

Begeleider	Tweede lezer
dr. Evert Haasdijk	dr. Fetsje Bijma

Fact. der Exacte Wetenschappen
Vrije Universiteit, Amsterdam

Stage uitgevoerd bij: Big Fellows, in opdracht van de Dienst Gezondheid Jeugd Zuid-Holland Zuid.

VOORWOORD

Deze scriptie is onderdeel van de onderzoeksstage van de Master Business Analytics aan de Vrije Universiteit te Amsterdam. De onderzoeksstage vormt de afsluiting van deze tweejarige Master waarin de studenten zich binnen meerdere disciplines ontwikkelen. De positie van mijn onderzoek binnen het curriculum is de "Computational Intelligence"-richting ("Data science").

Gedurende zes maanden heb ik namens mijn stagebedrijf (Big Fellows) de mogelijkheden van datagedreven, voornamelijk "Machine Learning", oplossingen binnen de VSV-problematiek verkend. Gezien de succesvolle toepassingen van "Machine Learning"-technieken binnen de meest uiteenlopende branches, (b)leek het een mooie kans om deze technieken op deze maatschappelijk relevante problematiek toe te passen. Dit stageonderzoek is tot stand gekomen in samenspraak met onze klant, de Dienst Gezondheid en Jeugd ZHZ.

Graag wil ik mijn dank uitspreken voor het enthousiasme waarmee de Dienst Gezondheid en Jeugd ZHZ dit onderzoek is ingegaan in de personen van Erwin Keuskamp en Ineke Noorthoek. Tevens wil ik mijn begeleiders vanuit Big Fellows, Martijn Minderhoud en Jaring Hiemstra, bedanken voor hun inzet om van een van theoretische scriptie tot een praktijkoplossing te komen. Als laatste wil ik uiteraard de begeleiding vanuit de VU bedanken. Allereerst Evert Haasdijk, met wie ik ook al mijn Research Paper succesvol heb doorlopen. Hij heeft mij wederom van waardevolle input voorzien. Ook de tweede lezer, Fetsje Bijma, wil ik bedanken voor het lezen van deze scriptie.

Anton Kuijer
Augustus 2015

INHOUDSOPGAVE

AFKORTINGEN

VSV Voortijdig Schoolverlaters / Voortijdig School Verlaten

DGJ (ZHZ) Dienst Gezondheid & Jeugd (Zuid-Holland Zuid)

GBA Gemeentelijke Basisadministratie

LBA Leerling-basisadministratie

DUO Dienst Uitvoering Onderwijs

BBL Beroeps Begeleidende Leerweg

BOL Beroeps Opleidende Leerweg

CBS Centraal Bureau voor de Statistiek

OCW Het ministerie van Onderwijs, Cultuur en Wetenschap

NJI Nederlands Jeugd Instituut

AOB Algemene Onderwijs Bond

JEDM Journal of Educational Data Mining

mbo Middelbaar beroepsonderwijs

vo Voortgezet onderwijs

CART Classification And Regression Tree

BDT Boosted Decision Tree

SVM Support Vector Machine

NN Neural Network

LR Logistische Regressie

RF Random Forest

ROC Receiver Operating Characteristic

Deel I

EERSTE FASE ONDERZOEK

1

INLEIDING

1.1 ACHTERGROND

De kenniseconomie die zowel op nationaal als Europees niveau wordt nagestreefd (2000, Strategie van Lissabon) is hoofdzakelijk afhankelijk van de kwaliteit van het onderwijs. Er kan gesteld worden, op basis van prestatie-indicatoren als schoolverlaten, slagingspercentages en Europese richtlijnen, dat het redelijk gesteld is met het onderwijs in Nederland. Door middel van intensieve begeleiding en invoering van de startkwalificatieplicht (2007) probeert men het aantal voortijdig schoolverlaters (VSV[1]) terug te dringen. Ondanks deze maatregelen, krijgt men moeilijk vat op de groep die er niet in slaagt een startkwalificatie te halen. De verantwoordelijkheid voor het aantal VSV ligt bij de regio's. Zij kampen met de moeilijke opdracht dit aantal terug te dringen.

1.2 PROBLEMATIEK

Ondanks het feit dat men steeds beter slaagt in het terugdringen van VSV, wil men dit aantal nog verder reduceren. De categorie ouder dan 18 jaar is hierbij specifiek interessant, mede omdat deze groep als een van de weinige is toegenomen tussen 2012 en 2013 (onderwijscijfers.nl). Schoolverlaters hebben meer moeite met het vinden van een baan en deze te behouden en er worden verbanden gesuggereerd tussen schoolverlaten en het in contact komen met justitie [1]. Ook de kenniseconomie die wordt nagestreefd gedijt niet bij een hoog VSV-percentage.

Naast de leerplicht is de startkwalificatieplicht (2007) ingevoerd waardoor leerlingen zonder een startkwalificatie verplicht zijn tot hun 18e jaar door te leren. De groep hardnekkige schoolverlaters blijft echter bestaan en hun handelen wordt vaak veroorzaakt door een complexe interactie tussen factoren. De regio's hebben de wens om door meer en beter inzicht effectiever en efficiënter te kunnen werken. Vandaar ook dat in de afgelopen jaren een preventieve aanpak hoger op de agenda is komen te staan. *Kan men schoolverlaten zien aankomen?*

De Dienst Gezondheid en Jeugd Zuid-Holland Zuid (DGJ ZHZ) bezit gegevens van alle (ex-)leerlingen (binnen de regio) tot en met hun 23e levensjaar. Indien nodig, worden deze leerlingen door middel van specifieke methodieken (gericht op individuen) begeleid. Niet iedereen kan begeleid worden. Een juiste aanwending van de (beperkte) middelen is daarom essentieel. In de huidige situatie krijgt elke VSV (18+ zonder startkwalificatie en niet ingeschreven aan een opleiding) in de 18 tot 23 leeftijdscategorie een brief. De

1 Leerlingen die het onderwijs zonder startkwalificatie (havo-, vwo- of mbo-niveau 2 en hoger diploma) verlaten.

jongeren die hier op reageren kunnen worden begeleid. Doordat de middelen beperkt zijn, kan het lonen de aanwending van de middelen (begeleiding) te verbeteren.

1.3 DOEL

Al het bovenstaande tezamen, suggereert dat "Machine Learning"-technieken hier mogelijk een rol spelen. Dit zou antwoord kunnen geven op de volgende vragen: *Wat gebeurt er met jongeren zonder startkwalificatie na hun 18e verjaardag? Wie haalt alsnog een startkwalificatie en wie niet en kan dit voorspeld worden? Is een beter gerichte begeleiding mogelijk?*

Het doel van dit onderzoek is om een eerste stap in de richting van datagedreven oplossingen te bieden. De complexe interactie tussen enkele kenmerken van succes kunnen mogelijk inzichtelijk worden gemaakt om zodoende een voorspelling te doen voor nieuwe gevallen. De mogelijkheden en toepassingen hiervan zijn divers. Zo kan een voorspellend model hulp bieden bij het gericht inzetten van begeleiding: het onderscheiden van "kansarmen" en "kansrijken". Dit onderzoek focust zich op "Machine Learning", om o.a. de samenhang van de factoren te onderzoeken.

Er is besloten dit onderzoek toe te spitsen op het mogelijk voorspellen van het al dan niet halen van een startkwalificatie na de 18e verjaardag en of dit, in combinatie met (ouderwetse) data analyse, inzicht kan bieden in (combinaties) van factoren die bijdragen aan het succes van de 18-plussers.

1.4 ONDERZOEKSVRAGEN

1.4.1 *Hoofdvraag*

Is het mogelijk om met behulp van "Machine Learning" te voorspellen welke 18-jarigen, zonder startkwalificatie, deze alsnog zullen halen in de daaropvolgende vijf jaar?

1.4.2 *Deelvragen*

- Kan met behulp van data-analyse en gesprekken met experts voldoende kennis opgedaan worden om een succesvol voorspellingsmodel in te richten?
- Welke kenmerken zijn in dit model belangrijk en hoe verhouden zij zich tot de in de literatuur gedefinieerde kenmerken?

1.5 BIG FELLOWS

Het stagebedrijf namens wie ik dit onderzoek uitvoer is Big Fellows, gevestigd te Utrecht. Deze start-up richt zich op het aanbieden van datagedreven oplossingen binnen de publieke sector. Met adviesbureau "Hiemstra & de Vries" als grote broer, probeert Big Fellows door middel van een combinatie van domeinkennis en kennis uit data, de manier van werken binnen de publieke sector te veranderen. Door afstudeerders met een technische achtergrond de kans te bieden hun onderzoek voor een echte klant uit te voeren,

bieden zij de klant mogelijkheden om op een betaalbare wijze inzicht te geven in de vele mogelijkheden van data-onderzoek.

1.6 STRUCTUUR

Deze scriptie begint met een bondig overzicht van zowel studies binnen het VSV domein als van Machine Learning toepassingen op onderwijs in ruime zin (Educational Data Mining). Met behulp van de literatuur kan een inschatting gemaakt worden van belangrijke kenmerken (ook wel *features* genoemd), te verwachten resultaten en praktische waarde van dit onderzoek. Na het literatuur hoofdstuk zullen de beschikbare data, en de mogelijkheden hiervan, besproken worden in relatie tot de uit de literatuur verkregen inzichten. Deze hoofdstukken zijn niet essentieel voor de aan tijd gebonden lezer.

Vervolgens zullen enkele voorbeelden uit het literatuurhoofdstuk getoetst worden door middel van data-analyse. In dat hoofdstuk zullen tevens vermoedens rond studiesucces van de DGJ ZHZ onderzocht worden om zo meer inzicht te krijgen in de problematiek.

Deze vermoedens (indien gegrond) worden vervolgens vertaald naar kenmerken die opgenomen kunnen worden in een model. In de modelleringshoofdstukken worden de kenmerken opgenomen in voorspellingsmodellen.

In de daaropvolgende hoofdstukken worden de uitkomsten van de modellen en de best presterende *feature subsets*, combinaties van kenmerken, besproken.

2

LITERATUUR ONDERZOEK: NEDERLAND

2.1 OPBOUW

Dit hoofdstuk biedt inzicht in de VSV problematiek en de oorzaken hiervan op basis van het literatuuronderzoek: *Kwalitatief onderzoek naar de risicofactoren van schoolverzuim en voortijdig schoolverlaten. Haverkort, Dordrecht, april 2014.* [2]
Haverkort heeft dit onderzoek in opdracht van de DGJ ZHZ uitgevoerd, net als deze scriptie. Tevens presenteert dit hoofdstuk een kort overzicht van VSV in Nederland en sluit af met een stukje over (statistische) data-analyse op dit gebied (in Nederland). De onderzoeken die hierbij gebruikt zijn, zijn geselecteerd op basis van hun overeenkomst met dit onderzoek en op basis van online vindbaarheid. Hierdoor kan het zijn dat enkele interessante onderzoeken op dit gebied ontbreken.

"Volgens een rapport van Allen en Meng (2010) zijn de drie belangrijkste redenen van schooluitval: verkeerde opleidingskeuze (27%), gezondheidsproblemen, zowel fysiek als psychisch (20%) en meteen willen werken (15%)" - Haverkort (2014).

2.2 HAVERKORT (2014)

Haverkort bespreekt in zijn scriptie 44 bekende probleemgebieden (VSV indicatoren) die worden bijgehouden in de leerling-basisadministratie (LBA). Het gaat hierbij voornamelijk om persoons-, achtergrond- of omgevingsfactoren waar de leerling zelf beperkt tot geen invloed op heeft.

Haverkort concludeert dat een eerder uitgebracht advies van een beleidsadviseur om het aantal probleemgebieden terug te dringen (van 44 naar 16) grotendeels moet worden overgenomen en reduceert dit aantal verder tot de negen voornaamste probleemgebieden. Tevens wordt er gesteld dat schoolverlaten zelden één oorzaak kent. Sterker nog, bij schoolverlaten is er sprake van een complexe interactie tussen de verschillende factoren (gecombineerde effecten). Desalniettemin worden beschermende factoren niet meegenomen in het onderzoek (ten behoeve van een overzichtelijk beeld).

Ook bevestigt Haverkort dat voortijdig schoolverlaters tegen allerlei problemen aanlopen en veroorzaken. Zo lijken er verbanden tussen schoolverlaten en het in contact komen met justitie, maar ook de kenniseconomie die wordt nagestreefd gedijt niet bij een hoog VSV-percentage. Ook hebben schoolverlaters meer moeite met het vinden van een baan en deze te behouden (Beckers en Traag, 2005, zoals beschreven in Haverkort). Kort (en simpel) gezegd: schoolverlaters kosten geld.

Voordelen van dalende schooluitval volgens Haverkort [2]:

- *Met een startkwalificatie staan jongeren sterker op de arbeidsmarkt. De werkloosheid, is twee keer lager onder jongeren met een startkwalificatie (OCW, 2013).*
- *De dalende schooluitval draagt bij aan het terugdringen van de jeugdwerkloosheid (OCW, 2013).*
- *Meer startkwalificaties betekent minder jeugdcriminaliteit. Voortijdig schoolverlaters zijn vijf keer vaker verdachte van een misdrijf (OCW, 2013).*
- *Meer startkwalificaties betekent ook lagere kosten in de zorg. Mensen met alleen basisschool of vmbo hebben hogere zorgkosten (COB, 2013).*

Over de complexe interactie van factoren, waarvan sprake is bij VSV, geeft Haverkort enkele voorbeelden. Allereerst de onderwijsondersteuning van de ouders. Volgens hem is dit vaak een beschermende factor, maar ook dikwijls een risicofactor. Ouders van VSV scoorden hoger op onderwijsondersteuning. Kennelijk zijn er omstandigheden waar onderwijsondersteuning een ander effect kan hebben (interactie tussen factoren). Een tweede voorbeeld dat wordt gegeven:

"Zo verlaten niet-westerse allochtonen weliswaar vaker het onderwijs zonder startkwalificatie, maar dit komt vooral door het gecombineerde effect van andere factoren, zoals leeftijd, het leeftijdsverschil met het gemiddelde van de klas, het aantal keren dat een leerling verdachte is geweest van een misdrijf, of het inkomstenniveau van het huishouden." (Pijpers, 2010) zoals geciteerd in Haverkort.

Het is overigens niet vreemd dat Haverkort zich beperkt tot één-op-één relaties, vooral gezien het feit dat een datagerichte aanpak binnen dit gebied (in Nederland) vrij nieuw is. Het onderzoek sluit af met een bespreking van de in zijn ogen negen voornaamste probleemgebieden (domeinen) van VSV en komt tot zes bepalende risicofactoren: ingrijpende levensgebeurtenissen, thuissituatie, opleiding, financiën/schulden, gezondheid en sociaal netwerk. Hierbij hebben bepaalde domeinen alleen of voornamelijk betrekking op de ouders.

2.3 VSV IN NEDERLAND: CIJFERS

Kennelijk zijn er een aantal probleemgebieden die een belangrijke rol spelen bij VSV. Maar hoe is het eigenlijk met VSV in Nederland gesteld? In het schooljaar 2013-2014 waren er 25.970 nieuwe voortijdig schoolverlaters (1,9% van alle leerlingen). Dat lijkt misschien veel, maar sinds 2004-2005 daalt het aantal VSV gestaag. Het dalingspercentage is ongeveer 0,3% per schooljaar. Allochtonen vallen verhoudingsgewijs vaker voortijdig uit dan autochtonen. Wel is het zo dat meer allochtone voortijdig schoolverlaters de weg terugvinden naar school. Ook vallen jongens vaker uit dan meisjes (2,4% ten opzichte van 1,5%). Dit valt op te maken uit een verslag van het Nederlands Jeugd Instituut (NJI) [3].

De meeste schoolverlaters (5,2%) zijn te vinden in het middelbaar beroepsonderwijs (mbo). Ter vergelijking, in het voortgezet onderwijs (vo) is dit slechts 0,5%. Binnen het mbo is de uitval het grootst in de eerste jaren. Van de mbo niveau 1 starters, verlaat meer dan een derde in het eerste schooljaar de school [3]. Sterker nog, volgens een artikel op de website van de Algemene Onderwijs Bond (AOB) [4], verlieten destijds vier op de tien mbo-leerlingen de school zonder diploma. Douma [4] citeert hierbij een onderzoek

van PricewaterhouseCoopers en het Kenniscentrum Beroepsonderwijs Arbeidsmarkt in opdracht van de mbo-raad. Volgens dit onderzoek haalde (rond 2007) slechts 61% het diploma. Dit percentage geeft mogelijk een wat vertekend beeld van de realiteit vanwege uitval door ziekte, psychische problemen, verhuizing of doorstroom tussen de mbo niveaus. Hier is helaas weinig zicht op en dit speelt waarschijnlijk op vergelijkbare wijze een rol bij het onderzoek in Zuid-Holland Zuid.

Volgens het NJI [3] is er bij leerlingen uit éénoudergezinnen sprake van een tweemaal zo hoog schoolverlaters percentage. Dit hangt duidelijk samen met een laag inkomen van de ouders. Er zijn wel veel allochtone leerlingen die uitvallen, maar er is geen direct verband tussen herkomst en voortijdig schoolverlaten. Een verklaring moet eerder worden gezocht in het gegeven dat zij vaker afkomstig zijn uit een gezin met een laag inkomen [3]. Een andere verklarende variabele voor schoolverlaten bleek het leeftijdsverschil met klasgenoten. Als dit verschil groot is, dan lopen de leerlingen een grotere kans om de school te verlaten.

2.4 VSV ONDERZOEK NOORDERPOORT

Er zijn dus behoorlijk wat gegevens beschikbaar over VSV in Nederland. Het is daarom interessant om te bekijken welke initiatieven hierbij een stap verder zijn gegaan dan de boven beschreven cijfers. Onderzoek Noorderpoort [5] (mbo-scholengemeenschap) is hier een voorbeeld van. Het onderzoek is uitgevoerd in samenwerking met ROC Nijmegen en ROC Twente met financiële en/of inhoudelijke bijdragen van Kennisnet, MBO15 en saMBO-ICT. De rapportage biedt handvatten en informatie bij het onderzoeken van datamogelijkheden bij het vergroten van studiesucces. Daarnaast worden de gepresenteerde aanpakken toegepast op de bijbehorende praktijkcase van Noorderpoort. Zo is er geprobeerd de Noorderpoort mbo-scholen in staat te stellen het studiesucces te vergroten *"op basis van voorspellende indicatoren in plaats van op basis van historische gegevens"*[5].

De discussie rond studiesucces wordt volgens de rapportage op basis van feiten, maar vaak ook op basis van overtuigingen en aannames gevoerd. Vandaar dat het objectieve karakter van een (correct) uitgevoerde data-analyse aansprak en tot de analyses leidde. De publicatie roept bestuurders en managers op om op zoek te gaan naar op feiten gebaseerde inzichten. Daarnaast vinden er, volgens de rapportage, ook discussies plaats over sturen op basis van voorspellende indicatoren in plaats van op basis van historische gegevens van een student. Al met al een interessante ontwikkeling waarbij men ook voornemens lijkt beter te willen voorspellen, maar dit lijkt ingestoken vanuit inzicht genererende data-analyse (in plaats van voorspellingen zoals bij "Machine Learning").

De praktijkcase richtte zich op het doorzoeken van gestructureerde data uit verscheidene databases om deze te analyseren (insight in plaats van foresight). Met als doel om objectieve feiten in het gesprek te brengen. Op basis van een literatuuronderzoek kwam men al tot de conclusie dat er meerdere belangrijke elementen zijn bij het bepalen van studiesucces. Twee voorbeelden hiervan zijn het switchen van opleiding tijdens de studieloopbaan en de verblijfsduur van de student op de instelling. Figuur 1 komt ook uit Kennisnet (2013) [5] en geeft weer welke factoren in de onderwijsliteratuur besproken worden als indicatoren van studiesucces. Dit zegt niet per definitie iets over de voorspel-

lende waarde van de kenmerken, het geeft slechts weer over welke onderwerpen er veel geschreven is.

Indicatoren ↓ / Publicaties →	Blauw (2009)	CPB (2012)	ECBO (2012)	Elffers (2011)	Heijden (2012)	Herweijer (2008)	Kappe (2011)	KBAN (2010)	Neuvel (2011)	Pellenaux (2003)	ResearchNed (2009)	ResearchNed (2010)	Ritzen (2008)	ROA (2009)	Rosenthal (1998)	Traag (2012)	Wolff (2010)
Geslacht student	•	•		•	•	•			•				•	•	•	•	
Leeftijd student (t.o.v. leerplichtleeftijd)	•	•			•										•		
Etniciteit (allochtoon / autochtoon)	•			•		•		•	•			•	•	•	•	•	•
Woonsituatie (1- of 2-oudergezin / uit- of thuiswonend)		•		•		•						•	•		•		
Cultureel kapitaal gezin/thuissituatie				•		•							•			•	•
Sociaaleconomische situatie buurt (ook urbanisatiegraad)						•								•	•	•	
Ziekte, handicap, stoornis (kan zijn LGF / 'rugzakje')						•						•					
IQ student							•						•				
Cognitieve vaardigheden van de leerling		•					•						•			•	
Financiële situatie / schulden van de student		•				•											
Student in aanraking geweest met politie/justitie		•														•	
Vooropleiding student (PO en VO, incl. CITO-score)	•	•	•			•		•	•		•	•				•	
Gemiddeld VO-cijfers (totaal en wiskunde)					•												
Verzuimgeschiedenis VO (spijbelen / schorsingen)						•								•			
Doublures in vooropleiding														•			
Rol van vrienden / klasgenoten (peers)										•		•	•			•	
Intrinsieke motivatie & zelfvertrouwen mbt keuze						•	•			•	•		•			•	
(On)zekerheid studiekeuze			•									•					
Opleidings- en beroepsbeeld ('binding')				•							•						
Niveau (1-4) en traject (BOL/BBL) MBO	•		•					•									
Datum aanmelding/inschrijving		•			•						•						
Bezoek aan voorlichtingsevenementen											•	•					
Welkomst- e/o intakegesprek aan de instelling											•	•					
Keuze voor instelling of voor stad/locatie											•	•					

Figuur 1: Publicaties over studiesucces en daarin onderzochte indicatoren. Bron: Noorderpoort onderzoek [5]

Het valt direct op dat etniciteits- en vooropleidingkenmerken veelbesproken zijn. Twee kenmerken waar in deze afstudeerscripie ook aandacht aan wordt besteed. Bij de praktijkcase van Noorderpoort zijn deze veelbesproken kenmerken uiteraard ook meegenomen. Zo bleek uit de data dat etniciteit bij het behalen van een diploma (ja of nee) nauwelijks een rol speelt. Dit in tegenstelling tot wat figuur 1 doet vermoeden. Geslacht daarentegen speelt wel een rol net als leeftijd (t.o.v. jaarlaag), aantal adressen gedurende studieloopbaan en betalingsachterstand (collegegeld). Een ander kenmerk dat een zeer belangrijke voorspeller bleek is het opleidingsniveau (mbo-niveau 1 t/m 4). Hoe hoger het niveau, des te groter de kans op het behalen van een diploma. Direct gerelateerd hieraan en minder voorspellend is de gekozen leerweg (BBL of BOL). Switchgedrag (wisselen van opleiding) lijkt echter weinig invloed te hebben. Op deze manier vormt het onderzoek een mooi fundament voor deze scriptie en zal er in komende hoofdstukken nog geregeld aan gerefereerd worden.

Een opmerkelijke kanttekening bij dit onderzoek is dat er wordt gesteld dat mensen met 50% "meer" vertraging een 3,7 keer grotere kans hebben op een diploma. Dit resultaat, naast het feit dat het contra-intuïtief lijkt dat vertraagden het beter doen dan niet vertraagden, doet wat vreemd aan aangezien in het onderzoek [5] de volgende definitie van vertraging wordt gehanteerd: *"Vertraging binnen de studieloopbaan is de proportie die een*

leerling langer of korter doet over de studie. De daadwerkelijke studieduur van de leerling is berekend op basis van de inschrijfdatum en de uitschrijfdatum van de studieloopbaan. De vertraging is berekend door de daadwerkelijke studieduur te delen door de nominale studieduur van een leerling." Deze definitie (*"proportie die een leerling langer of korter doet over de studie."*) impliceert in feite al dat het diploma behaald wordt. Want *"hoelang heb jij over je studie gedaan?"* suggereert: *"hoelang heeft het jou gekost je diploma te halen"*. Dus hier wordt in feite gesteld dat als je vertraging oploopt en dus langer over je studie doet (en dus meer jaren nodig hebt om je diploma te halen), dan is de kans groter dat je een diploma haalt. Mogelijk is dit resultaat simpelweg onhandig verwoord en moet er niet te veel gewicht aan toegekend worden. Dit resultaat is daarom moeilijk te duiden. Desalniettemin worden zoveel mogelijk resultaten van het onderzoek (Kennisnet, 2013) ook in dit onderzoek getoetst in de data-analyse voorafgaand aan de "Machine Learning".

Al met al doet Kennisnet [5] wat het belooft en dat is een aanpak presenteren voor dergelijk onderzoek en hier een praktijkvoorbeeld van geven in de vorm van Noorderpoort. Dit onderzoek is zeer bruikbaar voor deze afstudeerscriptie en kan bijvoorbeeld gebruikt worden om resultaten mee te vergelijken.

2.5 OVERIG

Afsluitend is er nog een onderzoek geweest in Nederland dat zeker de moeite waard is om te noemen. Jol et al. [6] hebben met behulp van logistische regressie bekeken in hoeverre het behalen van een startkwalificatie samenhangt met de gezinssituatie (insight). In het artikel staat de relatie tussen de gezinssituatie en het onderwijssucces van thuiswonende jongeren centraal. Een belangrijke uitkomst hiervan was dat het onderwijsniveau van de ouders een rol speelt bij het al dan niet behalen van een startkwalificatie. Maar ook de leeftijd van de moeder (hoe ouder hoe beter) en of de leerling opgroeit in een tweeoudergezin speelt een rol. Het verschil naar herkomst blijkt echter voornamelijk samen te hangen met andere kenmerken uit de gezinssituatie. De studie onderschrijft hiermee de eerder gepresenteerde resultaten van het Nederlands Jeugd Instituut [3].

2.6 CONCLUSIE

Dit hoofdstuk heeft de VSV-problematiek in Nederland uiteengezet en tevens bondig verkend welke Nederlandse onderzoeken van waarde zijn voor het onderzoek in Zuid-Holland. Belangrijke conclusies hierbij zijn dat VSV vooral op het mbo een probleem is. In Nederland wordt informatie hierover al heel lang in kaart gebracht, onder andere door het CBS. Het lijkt er echter op dat de ontwikkelingen zich langzaam in de richting van Machine Learning toepassingen begeven. Naast "inzicht" wordt namelijk ook de term "voorspellende waarde" steeds vaker gebruikt. Wat dat betreft past dit onderzoek goed binnen deze ontwikkeling. De inzichten verkregen uit de beschreven onderzoeken kunnen gebruikt worden om de modellen te verbeteren. Met andere woorden, in de *feature generation/selection*-fase kunnen deze onderzoeken geraadpleegd worden om tot de beste features te komen. Interessant hierbij is dat volgens veel van de onderzoeken, herkomst weinig voorspellende waarde heeft, maar vaak hand in hand gaat met andere (beter) voorspellende kenmerken zoals een lager inkomen.

De geraadpleegde onderzoeken suggereren dat er bepaalde achtergrondkenmerken in meer of mindere mate voorspellende waarde hebben. Studieresultaten worden in de onderzoeken nauwelijks besproken terwijl deze mogelijk ook een interessante indicator zijn van toekomstig studiesucces. Zeker omdat veel van de besproken kenmerken niet alleen tot uiting komen in het eindresultaat, maar natuurlijk ook in het traject (voorgaande schoolresultaten) daar naar toe. Denk hierbij bijvoorbeeld aan basisschoolresultaten die iets kunnen zeggen over iemands kansen op de middelbare school.

3

LITERATUURONDERZOEK: EDUCATIONAL DATA MINING

3.1 EDUCATIONAL DATA MINING

Dit onderzoek past (landelijk gezien) dus binnen de beweging waarbij onderwijs steeds meer wordt ondersteund met data. Internationaal gezien, past dit onderzoek echter ook goed binnen de ontwikkelingen van de laatste jaren waarbij "Machine learning"-technieken steeds breder worden ingezet, zo ook in het onderwijs. Men lijkt zich er steeds meer van bewust te zijn dat er winst te behalen valt door middel van efficiëntere begeleiding, beter aansluitende lesinvulling of een slimmere samenstelling van het vakkenpakket. Voorspellingsmodellen kunnen hier van waarde zijn.

Educational data mining is daar een mooi voorbeeld van. *The International Educational Data Mining Society* [7] tracht deze ontwikkeling te ondersteunen. Zij doen dit door middel van het organiseren van een jaarlijkse conferentie en de gratis uitgave van een vaktijdschrift (*Journal of Educational Data Mining, JEDM*). Dit online tijdschrift bevat uiteenlopende artikelen gebaseerd op verscheidene typen data. Voorbeelden van onderwerpen zijn: classificatie van relevante en irrelevante zinnen en woorden in een wiskundig probleem, inzicht in studentenresultaten, het analyseren van het taalgebruik van studenten, kenmerken van succes in onderwijskundige computerspelletjes inzichtelijk maken en vele andere onderwerpen.

Dit hoofdstuk bestaat uit een aantal onderzoeken (niet zozeer uit JEDM) waarbij "Machine Learning" is ingezet binnen onderwijs. Er is met name gezocht naar onderzoeken en toepassingen die qua opzet vergelijkbaar zijn met wat de Dienst Gezondheid en Jeugd voor ogen had bij aanvang van deze afstudeerstage. Ook hier geldt dat dit overzicht afhankelijk is van online vindbaarheid van de onderzoeken. Hierdoor kan het zijn dat enkele interessante onderzoeken op dit gebied ontbreken. Voor de geselecteerde onderzoeken is er bekeken hoe de diverse aanpakken presteerden om een reëel beeld te krijgen van de succeskansen van een voorspellingsmodel binnen onderwijs.

3.2 NOORWEGEN

Markussen et al. [8] hebben in Noorwegen een onderzoek uitgevoerd waarbij zij factoren die een rol spelen bij VSV in kaart brengen door middel van multinominale logistische regressie (afstudeerders, gezakten en schoolverlaters). Zij hebben hierbij een specifieke groep van bijna 10.000 leerlingen vijf jaar lang gevolgd op basis van *Public Register Data* (vergelijkbaar met GBA / LBA) en ingevulde survey's. Hun focus lag voornamelijk op het inzichtelijk maken van de problematiek. Op deze manier kwamen kenmerken als achter-

grond en leergedrag naar voren. Verreweg het meest voorspellend bleken eerder behaalde schoolresultaten. Het artikel beperkt zich tot het brengen van inzicht in tegenstelling tot het aanbieden van een *black-box algoritme* dat voorspelt. Deze meer klassieke statistische aanpak resulteert dus niet in een *performance measure* van het model door middel van *cross-validatie*, maar presenteert en verklaart het trainen van een model (vanuit een "Machine learning"-perspectief). Leo Breiman [9] (statisticus met veel publicaties rond *tree-based* algoritmes), onderschrijft dit door te stellen dat bij "Machine learning" de nadruk veel meer ligt op kwaliteit van voorspellingen (*predictive accuracy*) dan bij statistiek. Dit verklaart de twee stromen binnen dit type onderzoek.

Voor het VSV-onderzoek in Zuid-Holland Zuid zijn de data lang niet zo divers als bij Markussen et al. De uidaging zit hem hier in het behalen van een optimale voorspelling uit een beperkt aantal kenmerken. Ook ligt de focus voor deze scriptie meer bij het aanbieden van een voorspellingsmodel, om zo efficiëntere begeleiding mogelijk te maken, dan bij het bieden van inzicht.

3.3 SLOVENIË

In Slovenië is een vergelijkbaar onderzoek uitgevoerd [10] waarbij geprobeerd is het eindresultaat van een student op een bepaald type middelbare school te voorspellen op basis van de resultaten na één studiejaar. De data bij dit onderzoek bestond onder andere uit: persoonlijke en demografische data, basisschool prestaties en cijfers uit het eerste jaar middelbaar onderwijs. Met behulp van onder andere *decision trees* kwamen zij niet veel verder dan 60% accuracy. In het artikel wordt nog wel type onderwijs opgesplitst om voor bepaalde groepen tot betere voorspellingen te komen (69.7%). De kwaliteit van dit artikel, tevens hoofdstuk van een boek [10] (à $149), trek ik in twijfel aangezien er ook resultaten gepresenteerd worden waarbij een classifier getraind wordt om na vier jaar onderwijs (op een vier jaar durende opleiding) onderscheid te maken tussen: stopt onderwijs na eerste jaar, stopt onderwijs na eerste jaar of later, leerling met vertraging, studeert af met een A of B, studeert af met een C of D. Een deel van deze informatie is natuurlijk op basis van de data (of de afwezigheid hiervan) te bepalen, daar is geen classifier voor nodig. De onderzoekers nuanceren dit resultaat wel, maar staan naar mijn mening niet voldoende stil bij de implicaties hiervan. Desalniettemin een interessant initiatief om met "Machine learning" de begeleiding (*counseling*) op middelbare scholen te verbeteren.

3.4 NIEUW-ZEELAND

Kovačić [11] heeft in Nieuw-Zeeland onderzocht of met behulp van socio-demografische data (leeftijd, geslacht, afkomst, onderwijs, werk en invaliditeit) en studie omgeving (lesprogramma, semester) succesvolle en minder succesvolle leerlingen te onderscheiden zijn. Het gaat hier om een "small data" vraagstuk waarbij gedurende 3 jaar data is verzameld van 450 leerlingen. Het onderzoek probeert om met behulp van *decision trees* voorspellingen te doen en inzicht te bieden in de bepalende kenmerken.

De afhankelijke variabele in dit onderzoek is binair (**Pass:** succesvolle afronding of **Fail:** onsuccesvolle afronding dan wel terugtrekking of schoolwissel). Interessant hierbij is dat van school wisselende (waaronder verhuizende) leerlingen mogelijk wat ruis veroorzaken

in de data. Kovačić stelt dat een mogelijke oplossing hiervoor kan zijn om de afhankelijke variabele verder op te splitsen en een categorie van *voluntary transfer* en *withdrawals* toe te voegen. Bij het VSV-vraagstuk in Zuid-Holland Zuid valt echter niet af te leiden of mensen stoppen met onderwijs of bijvoorbeeld verhuizen en verder studeren.

Kovačić beste voorspelling wordt behaald door middel van *Classification And Regression Trees* (CART) en bedraagt 60.5% aan goed voorspelde gevallen. Om één en ander in perspectief te plaatsen: 52% van de leerlingen uit de dataset eindigt met een onsuccesvolle schoolafronding. Dit kan gezien worden als een *predict majority class benchmark*. Met andere woorden, als iedereen als onsuccesvol voorspeld wordt dan is de nauwkeurigheid (percentage juiste voorspellingen): 52%.

Deze score suggereert volgens Kovačić dat de verzamelde achtergrondkenmerken toch niet voldoende informatie bevatten en dat in het verleden behaalde onderwijsresultaten mogelijk interessanter zijn. Verder beschrijft Kovačić nog een onderzoek van Kember [12] met als belangrijkste conclusie dat achtergrondkenmerken minder goede voorspellers blijken dan gedacht, omdat zij slechts een startpunt zijn. Er zijn veel meer andere factoren die bijdragen aan de moeilijkheden die studenten ondervinden tijdens hun studie. Hij stelt vervolgens dat de in zijn onderzoek geleverde modellen, nog niet goed genoeg onderscheid maken tussen succesvolle en onsuccesvolle leerlingen.

Wat dat betreft zijn er goede mogelijkheden voor het onderzoek in Zuid-Holland Zuid. De beschikbare DUO-data (gehele onderwijsloopbaan) maken het vraagstuk ondanks het beperkte aantal socio-demografische kenmerken (geslacht, leeftijd etc.) mogelijk kansrijk. Ook het feit dat het aantal onderzochte leerlingen in Kovačić onderzoek beperkt is, bestempelt hij als mogelijke oorzaak van zijn niet optimale resultaten. In Zuid-Holland Zuid is een dataset van ruim 23.000 leerlingen beschikbaar dus daar zal dit probleem minder spelen.

3.5 PORTUGAL

Een ander in potentie interessant onderzoek op dit gebied heeft in Portugal plaatsgevonden. Cortez en Silva [13] hebben hier door middel van vragen- en cijferlijsten geprobeerd succesvolle- en niet-succesvolle studenten te onderscheiden. Deze vragenlijsten waren zeer uitgebreid en bevatten vragen over reistijd naar school, relatiestatus, werk en salaris van ouders, vrijetijdsbesteding, alcoholgebruik en ga zo maar door. Op deze manier is er een zeer compleet beeld van de leerlingen opgebouwd. Het doel was om met deze kenmerken het succes (ja of nee) te voorspellen aan het einde van periode drie en later om ook het cijfer te voorspellen. Om dit te doen zijn er diverse experimenten uitgevoerd met verschillende modellen en *features*. Zo zijn er modellen gebouwd die ook de cijfers uit periode één (P1) en twee (P2) meenemen om een inschatting te maken van het succes (of cijfers) in periode drie. Logischerwijs scoort het model met binaire response variabele (succes of niet in P3) met toegevoegde *features*, over de cijfers in P1 en P2, het best. De praktijkwaarde hiervan is beperkt, omdat het voorspellingsmoment, na periode twee, rijkelijk laat is om nog in te kunnen grijpen.

Het is goed om na te gaan, dat in dit geval er een voorspelling wordt gedaan of de leerling over gaat op basis van alle cijfers behalve die van de laatste periode. Dan is het natuurlijk niet vreemd dat er een hoge nauwkeurigheid behaald wordt (90%+). Sterker

nog, Cortez en Silva [13] gebruiken naast de cijfers uit P1 en P2 ook achtergrondkenmerken uit de vragenlijsten en bespreken de toegevoegde waarde hiervan nagenoeg niet. Aangezien de gebruikte data publiekelijk beschikbaar is, ben ik zo vrij geweest om dezelfde voorspelling te doen (succes of geen succes), maar dan met alleen de P1 en P2 cijfer data. Door gebruik te maken van een *two-class decision forest* heb ik de scores uit het onderzoek (binaire respons, P1 en P2 cijfers data) minimaal geëvenaard. Cortez en Silva ontkennen overigens niet dat de schoolresultaten het belangrijkst zijn, maar nemen nog wel een enkel achtergrondkenmerk in het model op, terwijl dit mijns inziens niet nodig is. Het is belangrijk te benadrukken hoe waardevol de schoolresultaten uit P1 en P2 zijn om P3 te voorspellen.

Voor deze scriptie is de paragraaf waar Cortez en Silva de resultaten naar aanleiding van een model, slechts op basis van persoonskenmerken, bespreken bijzonder interessant. Deze aanpak is interessant, omdat de waarde van achtergrondkenmerken met wisselende resultaten uit de literatuur naar voren komt. Helaas blijkt een experiment met achtergrondkenmerken slecht te werken. De resulterende *accuracy* (nauwkeurigheid) verslaat de *predict majority class benchmark* (voorspel dat iedereen slaagt) nagenoeg niet. Met andere woorden: het lijkt er op dat deze resultaten weinig concrete praktijkwaarde hebben.

In het onderzoek worden ook modellen getraind die het exacte eindcijfer voorspellen en een classificatie model die de cijfers onderverdeeld in vijf klassen. Dit resulteert in veel lagere scores en al helemaal als de cijfers uit P1 en P2 buiten beschouwing worden gelaten. Omdat die modellen voor deze scriptie minder relevant zijn, heb ik mij hierop minder gefocust en sluit ik niet uit dat deze in de praktijk wel waarde kunnen genereren. Tevens presenteren Cortez en Silva nauwkeurigheid als gebruikte *performance measure*. Daar ik geen zicht heb op de rest van de resultaten (overige mogelijke *performance measures*), omdat deze niet gepresenteerd worden in het artikel, kan er ook niet uitgesloten worden dat de modellen op andere criteria misschien beter scoren dan bovenbeschreven benchmark. Daarbij geeft het onderzoek inzicht in welke achtergrondfactoren een belangrijke, en welke een minder belangrijke, rol spelen.

In feite bevestigt dit onderzoek wederom dat studieloopbaan data, in dit geval in de vorm van cijfers, zeer veel voorspellende waarde heeft. Als inzicht minder belangrijk is en de focus een zo goed mogelijk voorspellingsmodel is (insight vs. foresight) komen de persoonskenmerken die van invloed zijn (bijvoorbeeld alcoholgebruik) vanzelf naar voren in iemands schoolresultaten. Daarom blijkt uit dit onderzoek dat studieresultaten de beste voorspellers zijn.

3.6 OVERIG

Op deze, of soortgelijke, manier zijn er nog meer onderzoeken gedaan. Kovačić [11] beschrijft op bondige wijze nog enkele interessante onderzoeken binnen dit gebied. Hieronder een kleine selectie van een aantal uitkomsten van andere, voor deze scriptie minder interessante, onderzoeken.

Zo noemt Kovačić het onderzoek van Woodman [14] die met behulp van binaire logistische regressie inzichtelijk maakte welke factoren van invloed zijn op slagen en zakken of het niet maken van een tentamen. Verklarende variabelen bleken onder andere: aantal behaalde wiskunde toetsen, niveau van het vak, aantal studiepunten e.d.

Simpson [15] gebruikte vergelijkbare methodieken om met registratiedata van vakken te bepalen dat het niveau van het vak, voorgaand onderwijs, vakkenpakket, socio-economische status, geslacht en leeftijd gezamelijk de succeskansen bepalen.

Kotsiantis et al. [16] gebruikten ook demografische variabelen en cijfers in meerdere "supervised learning"-algoritmes om studentensucces te voorspellen. Waar zij met alleen de demografische variabelen rond 65% goed voorspelden, resulteerde het toevoegen van de overige variabelen tot verbeterde resultaten.

Vandamme et al. [17] gebruiken *decision trees, neural networks* en *linear discriminant analysis* om risicogroepen van leerlingen te onderscheiden (low, medium en high risk). Zowel demografische als academische data zijn hiervoor gebruikt. Dit onderzoek is overigens vergelijkbaar met een groot deel van de eerder beschreven onderzoeken. Uiteindelijk bleken achtergrondkenmerken als geslacht, opleiding *en* beroep ouders en burgerlijke staat een lage voorspellende waarde te hebben. Het best scorende model kwam tot een classificatie nauwkeurigheid van 57%.

Als laatste voorbeeld hebben Yu et al. [18] zittenblijvende *freshmen* van Arizona State University geprobeerd te voorspellen met behulp van *decision trees*. Ook hier bleken geslacht en afkomst irrelevant.

Herrera [19] verklaart de verschillen in de bovenstaande uitkomsten (bijvoorbeeld het gebruik van demografische data versus studieresultaten) door te stellen dat wanneer een kenmerk voorspellend is, afhangt van de context. Niveau, opleiding, schoollocatie en dergelijke kunnen alle van invloed zijn. Er wordt dan ook gesuggereerd dat een uniek model per academisch niveau zeer interessant kan zijn.

3.7 CONCLUSIE

In dit hoofdstuk zijn de meest interessante buitenlandse onderzoeken in relatie tot de VSV-vraagstukken in Zuid-Holland Zuid beschreven. De algehele tendens is dat er aan de hand van onderwijsdata tot nog toe geen goede voorspellingen zijn gedaan. Dit heeft alles te maken met het feit dat de gebruikte technieken vrij nieuw zijn binnen dit vakgebied. Dit viel ook op te maken uit het feit dat enkele onderzoeken inhoudelijk sterk zijn opgezet, maar er technisch gezien, qua modellen, nog makkelijk winst te boeken valt. Zo is er een aantal onderzoeken geweest die zich volledig focussen op *decision trees* terwijl er een zeer breed scala aan modellen is zodat het testen van slechts één of twee modellen eigenlijk niet volstaat. Andere onderzoeken zijn aan de technische kant weer goed opgezet, maar lijken zich niet voldoende bewust van implicaties en praktische waarde. Het eerder beschreven voorspellen van "pass" of "fail" in periode 3 op basis van de cijfers aan het eind van periode 2 is hier een voorbeeld van.

Opvallend is dat studieresultaten vrijwel altijd goede indicatoren zijn. Aan de andere kant hebben achtergrondkenmerken wisselend voorspellende waarde. Dit in tegenstelling tot de meer theoretische studies die zijn gedaan (hoofdstuk over Nederlandse literatuur). De lagere voorspellende waarde van achtergrondkenmerken heeft meerdere oorzaken. Een veelgenoemd voorbeeld hiervan gaat over de context (locatie, onderwijssysteem, welvaart etc.). Vandaar ook dat de besproken "Machine learning"-resultaten geen garanties bieden voor de resultaten van dergelijk onderzoek op Nederlandse data. Wel is het zo dat voor het onderzoek voor de Dienst Gezondheid en Jeugd de data hoofdzakelijk uit school-

loopbaan kenmerken bestaan en dat de hoeveelheid data ook nog eens groter is dan bij de besproken onderzoeken. De tweede reden dat achtergrondkenmerken het slechter doen dan verwacht, komt waarschijnlijk doordat er een te grote verscheidenheid aan factoren is die invloed hebben op schoolsucces en VSV. Zo sprak Haverkort al over 44 probleemgebieden die ook nog eens, al dan niet gelijk, optreden. De meeste van deze factoren komen natuurlijk al tot uiting in de tot dan toe behaalde studieresultaten. Op het moment van voorspellen lijken behaalde studieresultaten daarom betere voorspellers.

Al met al kan gesteld worden dat er al een aantal keer eerder geprobeerd is studenten succes te voorspellen. Er zijn echter weinig oplossingen geboden die direct praktische waarde hadden. Met het onderzoek in Zuid-Holland Zuid hoop ik de toegevoegde waarde van "Machine Learning" bij de beschreven problematiek aan te tonen.

BIBLIOGRAFIE

[1] Tanja Traag en Olivier Marie. Voortijdig schoolverlaten, werkloosheid en delinquentie: cumulatie van risicogedrag onder jongeren in nederland. 2011.

[2] Wouter Haverkort. Literatuuronderzoek: Kwalitatief onderzoek naar de risicofactoren van schoolverzuim en voortijdig schoolverlaten. *Dienst Gezondheid Jeugd Zuid-Holland Zuid (Dordrecht)*, 2014.

[3] Nederlands Jeugd Instituut. Cijfers voortijdig schoolverlaten, 2014.

[4] L. Douma. Veel uitvallers op mbo, 2007.

[5] Big data, van hype naar actie: Op zoek naar waardevolle inzichten voor het vergroten van studiesucces. *Noorderpoort, Kennisnet, saMBO-ICT en MBO15*, 2013.

[6] Francis van der Mooren Christine Jol, Godelief Mars. Niet behalen startkwalificatie hangt samen met gezinssituatie. 2012.

[7] International Educational Data Mining Society. Journal of educational data mining.

[8] Nina Sandberg Eifred Markussen, Mari W. Frseth. Reaching for the unreachable: Identifying factors predicting early school leaving and non-completion in norwegian upper secondary education. *Scandinavian Journal of Educational Research*, 55(3):225–253, 2011.

[9] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci.*, 16(3):199–231, 2001.

[10] Vladislav Rajkovic Silvana Gasar, Marko Bohanec. *A Combined Data Mining and Decision Support Approach to Educational Planning*. Springer, 2003.

[11] Zlatko J. Kovačić. Early prediction of student success: Mining students enrolment data). 2010.

[12] D. Kember. Open learning courses for adults: A model of student progress. 1995.

[13] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. *In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*, 2008.

[14] R. Woodman. Investigation of factors that influence student retention and success rate on open university courses in the east anglia region. 2001.

[15] O. Simpson. Predicting student success in open and distance learning. *Open learning*, 21(2):125–138, 2006.

[16] Pierrakeas C Pintelas P. Kotsiantis, S. Predicting students' performance in distance learning using machine learning techniques. 2004.

[17] Meskens N. Superby J.-F. Vandamme, J.-P. Predicting academic performance by data mining methodsn. *Education Economics*, 15(4):405–419, 2007.

[18] DiGangi S. Jannasch-Pennell A.-Lo W. Kaprolet C. Yu, C. H. A data-mining approach to differentiate predictors of retention. 2007.

[19] O. L. Herrera. Investigation of the role of pre- and post-admission variables in undergraduate institutional persistence, using a markov student flow model. 2006.